Aus dem Institut für Molekularbiologie und Tumorforschung
der Philipps-Universität Marburg
Geschäftsführender Direktor: Prof. Dr. R. Müller

Statistische Auswertung und Interpretation von hochdimensionalen molekularbiologischen Datensätzen

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Humanbiologie
(Dr. rer. physiol.)

dem Fachbereich Medizin der

Philipps Universität Marburg

vorgelegt von

Birgit Samans
aus Aachen

Marburg, 2008

Angenommen vom Fachbereich Medizin der Philipps-Universität Marburg
am 14.10.2008

Gedruckt mit der Genehmigung des Fachbereiches.

Dekan: Prof. Dr. Matthias Rothmund
Referent: Prof. Dr. Martin Eilers
Koreferent: Prof. Dr. Karl-Heinz Grzeschik

Inhaltsverzeichnis

Abkürzungsverzeichnis

Sofern nicht hier aufgeführt, wurden Abkürzungen entsprechend den Maßangaben der IUPAC (*International union of pure and applied chemistry*) und denen des SI-Systems (*System Internationale de l´Unité*) verwendet. Spezielle Abkürzungen für Fachtermini, die nicht in dieser Liste enthalten sind, werden jeweils im Text erläutert.

4-OHT	4-Hydroxytamoxifen
A	Adenin
Abb.	Abbildung
ALV	avian leukemia virus
bp	Basenpaare
bzw.	Beziehungsweise
C	Cytosin
ca.	Zirka
cDNA	complementary DNA
CSHLmpd	Cold Spring Harbor Laboratory Mammalian Promotor Database
Cy3	Cyanin 3
Cy5	Cyanin 5
d.h.	das heißt
DAVID	Database for Annotation, Visualization and Integrated Discovery
DBTSS	DataBase of Transcriptional Start Sites
DNA	Desoxyribonukleinsäure
dsRNA	doppelsträngige Ribonukleinsäure
E. coli	Escherichia coli
EASE	Expression Analysis Systematic Explorer
EC	Enzyme Commission
ER	Estrogen-responsiven
et al.	und andere
FDR	false discovery rate
G	Guanin
GCM	Global Cancer Map
GNF2	Novartis Research Foundation SymAtlas
GSEA	Gene Set Enrichment Analysis
HTML	Hypertext Markup Language
HTS	High throughput screen
IMT	Institut für Molekularbiologie und Tumorforschung
KEGG	Kyoto Encyclopedia of Genes and Genomes
log	Logarithmus
max	Maximum
min	Minimum

miRNA	mikro Ribonukleinsäure
mRNA	messenger Ribonukleinsäure
MW	Mittelwert
mya	million years ago
neg.	negative
NES	normalised enrichment score
NIAID	National Institute of Allergy and Infectious Diseases
PCR	Polymerase Chain Reaction
PET	'paired-end ditag'
PFAM	Protein Families
PIR	Protein Information Resource
pos.	Positive
PSSM	Positions-specific scoring matrices
PWM	Position weight matrices
RISC	RNA-induced silencing complex
RNA	Ribonukleinsäure
RNAi	RNA-Interferenz
rRNA	ribosomale Ribonukleinsäure
RT/PCR	reverse transcription-polymerase chain reaction
s.	Siehe
SAM	significance analysis of microarrays
SD	Standard deviation
shRNA	short hairpin Ribonukleinsäure
siRNA	small interference Ribonukleinsäure
snoRNA	Small nucleolar Ribonukleinsäure
T	Thymin
Tab.	Tabelle
TBP	TATA-Box bindende Protein
TFBS	Transcription factor binding site
TSS	Transcription start site
u.a.	unter anderem
UTR	untranslated region
z.B.	zum Beispiel
z.T.	zum Teil

1 Einleitung

Der Einsatz von Hochdurchsatz-Methoden ist in der molekularbiologischen Forschung zu einem elementaren Werkzeug geworden. Während man in der klassischen Molekularbiologie versucht, anhand einzelner Aspekte zelluläre Prozesse zu analysieren, zielen die Hochdurchsatz-Methoden auf die Betrachtung mehrerer Faktoren zum gleichen Zeitpunkt, um so eine globale Sicht der dynamischen Veränderungen in der Zelle zu erhalten. Eine der ersten Einsatzgebiete von Hochdurchsatz-Technologien war die Entwicklung der automatisierten DNA-Sequenzierung. Diese ermöglichte 1995 erstmals die vollständige Sequenzierung des Genoms des gram-negativen Bakteriums *Haemophilus influenzae* (Fleischmann et al. 1995). Inzwischen liegen mehr als zweihundert sequenzierte Genome in Datenbanken vor. Hierzu gehört auch das humane Genom, dessen vollständige Sequenzierung im Jahre 2001 im Rahmen des *Human Genom Project* abgeschlossen wurde (Lander, Linton et al. 2001; Venter, Adams et al. 2001).

Von der Sequenzierung des humanen Genoms hatte man sich erhofft, Gene zu identifizieren, die an der Entstehung von Krankheiten beteiligt sind, und so zielgerichtet Ansätze für die Vorbeugung, Diagnose und Behandlung verschiedenster Erkrankungen zu erhalten. Diese Erwartungen konnten jedoch nicht erfüllt werden. Es wurde schnell deutlich, dass viele Krankheiten wie z.B. Herz-Kreislauf-, Krebserkrankungen oder die Alzheimer-Erkrankung nicht durch einfache Ursache-Wirkungsbeziehungen erklärt werden können, da sich die Mechanismen innerhalb der Zelle als deutlich komplexer erwiesen, als ursprünglich angenommen. So wurden z.B. verschiedene Gene identifiziert, die an der Entstehung der Alzheimer-Krankheit beteiligt sind, diese stellen jedoch Bestandteile eines komplexen Netzwerks molekularer und struktureller Komponenten dar, welches zudem noch durch Umweltfaktoren beeinflusst wird, wodurch das Verständnis der Entstehung der Krankheit weiterhin erschwert wird.

Eine wichtige Erkenntnis der Sequenzierung des humanen Genoms ist, dass die Anzahl der Gene mit 25.000 – 30.000 deutlich geringer ist als bisher angenommen und damit einer ähnlichen Anzahl entspricht, wie sie auch bei anderen Vertebraten zu finden ist (Lander, Linton et al. 2001; Venter, Adams et al. 2001; Aparicio 2002; Waterston,

Lindblad-Toh et al. 2002). Da die Anzahl der Gene bis zu diesem Zeitpunkt als Schlüsselelement für die höhere Entwicklung angenommen worden war, stellte sich nun die Frage, welche Prozesse stattdessen für die Komplexität höherer Organismen relevant sind. Hier wurde u.a. die Bedeutung von nicht-kodierenden RNA-Sequenzen erkannt. Der Vergleich verschiedener Organismen ergab, dass ihr Anteil mit zunehmender Komplexität der Organismen zunimmt. Inzwischen konnte es gezeigt werden, dass nicht-kodierende RNA Sequenzen eine wichtige Rolle auf verschiedenen Ebenen der Regulation übernehmen (Chu and Rana 2007).
Neben den nicht-kodierenden RNA-Sequenzen sind mittlerweile eine Reihe von weiteren Faktoren bekannt, z.B. Histon-Modifikation, Methylierung und Ubiquitinierung, die ebenfalls von Bedeutung für die Regulation zellulärer Vorgänge sind. Zelluläre Mechanismen können somit nicht durch einfache Ursache-Wirkungsbeziehungen beschrieben werden, sie stellen eher ein komplexes Netzwerk an Prozessen dar. Zur Untersuchung der Funktionen von Genen und Genprodukten bzw. von regulatorischen Prozessen wurden deshalb verschiedene Hochdurchsatz-Methoden etabliert. Hierzu gehören auch die in dieser Arbeit beschriebenen Methoden, die DNA-Chiptechnologie, die miRNA-Chiptechnologie und das RNAi-Screening-Verfahren. Diese drei Plattformen untersuchen verschiedene Ebenen der zellulären Prozesse. Während die DNA-Chiptechnologie mit dem Vergleich der mRNA-Level in verschiedenen Phänotypen die Transkriptionsebene untersucht, betrachtet die miRNA-Chiptechnologie die Regulation auf Translationsebene. Das RNAi-Screening untersucht den Phänotyp einer Zelle nach dem seriellen Ausschalten einzelner Gene und damit die Regulation auf Proteinebene.

Für die Etablierung von Hochdurchsatz-Methoden waren einige Voraussetzungen notwendig. Hierzu gehören, neben den Sequenzinformationen, automatisierte und standardisierte Produktionsbedingungen, die eine Parallelisierung und Miniaturisierung der Assayformate ermöglichen, effizient arbeitende Bildanalyseprogramme bzw. ausreichende Computerkapazitäten zur Analyse von hochdimensionalen Datensätzen. Zudem sind Algorithmen notwendig, deren Ziel es ist die Daten auszuwerten und zu interpretieren, um so Gesetzmäßigkeiten zu Erkennen und diese in den Kontext biologischer Vorgänge zu setzen.

Hochdimensionale Daten beinhalten einige Herausforderungen. So steht einer Vielzahl von getesteten Genen häufig nur eine geringe Anzahl von Proben gegenüber. Dies ist eine Situation, die bei statistischen Auswertungen sonst eher umgekehrt auftritt. Dementsprechend sind die bereits vorhandenen statistischen Methoden nur zum Teil für die Auswertung der Daten geeignet. Bei der Durchführung eines Hochdurchsatz-Experimentes ist es zudem schwierig, die Qualität aller Tests gleichermaßen zu kontrollieren, was sich häufig durch ein sehr hohes Rauschen der Daten zeigt.

Das Fehlen von standardisierten Qualitätskriterien und Datenanalysemethoden hat sich in der Vergangenheit als eines der Probleme für die effiziente Implementierung von Hochdurchsatz-Methoden erwiesen (Kaul 2005). Die Etablierung einer standardisierten Auswertungsroutine ist somit von großer Bedeutung für eine erfolgreiche Durchführung der Experimente.

Die Datenanalyse beinhaltet folgende Schritte

- Experimentelles Design
- Qualitätskontrolle
- Normalisierung
- Selektion von Hits
- Identifizierung von Mustern innerhalb der Ergebnisliste, die dann in den Kontext mit biologischen Informationen gestellt werden sollen

Im Rahmen dieser Arbeit wird die statistische bzw. bioinformatische Auswertung der Datensätze verschiedener Hochdurchsatz-Plattformen beschrieben.
Für die DNA-Chiptechnologie, die bereits seit Anfang der 90er Jahre Anwendung in der Untersuchung der Genexpression findet, sind bereits standardisierte Methoden zur Auswertung der Daten etabliert worden. Diese Methoden werden in Kapitel 2 dargestellt. Da sich in der Vergangenheit gezeigt hat, dass die Interpretation der Daten im Kontext der biologischen Fragestellung sich trotz effizienter Datenanalyse als schwierig gestaltet, wurde in den letzten Jahren stärker darauf fokussiert, bioinformatische Methoden zur funktionellen Interpretation der Daten zu etablieren. In Kapitel 3 werden einige dieser Methoden dargestellt und anschließend auf einen Datensatz angewendet. Ein Aspekt der funktionellen Analyse ist die Untersuchung von

ko-regulierten Genen eines Datensatzes auf eine mögliche Regulation durch einen gemeinsamen Transkriptionsfaktor. Eine Methode, die diese Fragestellung untersucht und im Rahmen dieser Arbeit entwickelt wurde, wird in Kapitel 4 dargestellt.
In Kapitel 5 wird dann die Etablierung der statistischen Auswertung von miRNA-Mikroarrays beschrieben. Die methodischen Grundlagen bilden hierfür die Verfahren, die auch zur Analyse von cDNA-Mikroarray-Plattformen verwendet werden (s. Kapitel 5). Im letzten Kapitel wird dann die Etablierung der statistischen Auswertung der RNAi-Screening-Methode aufgezeigt.

2 Methoden zur statistische Auswertung von cDNA-Mikroarrayversuchen

2.1 Programme

R/Bioconductor

Bioconductor ist eine *open source* Softwareprojekt, basierend auf der Skriptsprache R, welches statistische und graphische Methoden für die Analyse genetischer Daten beinhaltet (www.bioconductor.org). Es stehen verschiedene Module zur Verfügung, die speziell für die Auswertung von Mikroarray-Experimenten entwickelt wurden. Bei den im Folgenden beschriebenen Methoden zur Auswertung von Mikroarray-Experimenten wurden die Packete *limma* und *multtest* verwendet. Diese Skripte wurden erweitert, so dass die beschriebene Auswertungsroutine standardisiert und automatisiert durchgeführt werden kann.

2.2 DNA-Mikroarrays

Es werden hauptsächlich zwei verschiedene Arten von DNA-Mikroarrays verwendet. Einerseits cDNA-Mikroarrays, bei denen cDNA-Sequenzen auf die Oberfläche eines Objektträgers geprintet werden und anderseits sogenannte *GeneChips*® der Firma Affymetrix, die auf synthetisch hergestellten Oligonukleotiden beruhen. Da am Institut für Molekularbiologie und Tumorforschung (IMT) vorwiegend mit cDNA-Mikroarrays gearbeitet wird, beziehen sich die im Folgenden dargestellten Methoden auf diese Technologie.

2.2.1 Experimentelle Durchführung eines cDNA-Mikroarrayversuches

Die experimentelle Durchführung des Mikroarrayversuches ist in Abbildung 1 dargestellt. Die mRNA wird aus den zu untersuchenden Proben extrahiert und mit der *T7-RNA*-Polymerase amplifiziert. Die so entstandene mRNA wird anschließend in cDNA umgeschrieben und im gleichen Schritt wird jeweils eine Probe mit dem Fluoreszenzfarbstoff Cyanin-3 und die andere Probe mit Cyanin-5 markiert. Die beiden

unterschiedlich markierten cDNA-Proben werden dann auf einem cDNA-Mikroarray hybridisiert. Hierbei binden die markierten cDNA-Sequenzen an ihren komplementären Gegenpart auf dem Array. Anschließend wird das Fluoreszenzsignal der beiden Farbstoffe von jedem Spot des DNA-Mikroarrays mittels eines Laserscanners ausgelesen und in Signalintensitäten umgerechnet.

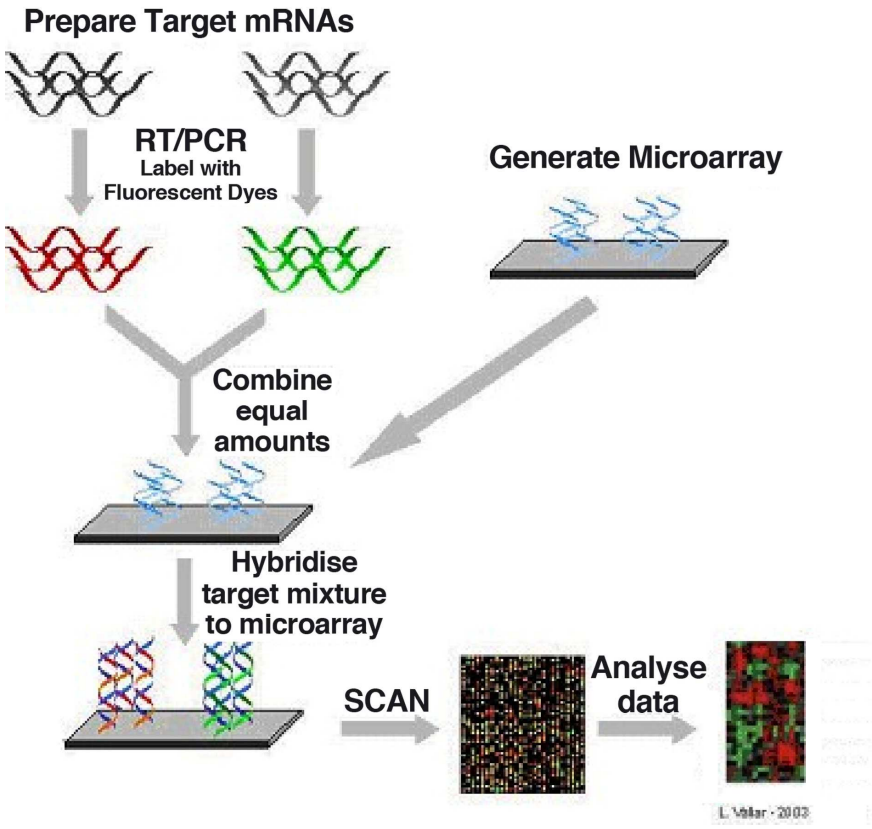


Abb. 1: Experimentelle Durchführung eines Mikroarrayversuches. Die aus den Proben extrahierte mRNA wird mit der *T7-RNA*-Polymerase amplifiziert. Die daraus resultierende mRNA wird dann in cDNA umgeschrieben. Während dieses Schrittes wird die eine Probe mit dem Fluoreszenzfarbstoff Cyanin-3 und die andere Probe mit Cyanin-5 markiert. Diese beiden unterschiedlich markierten cDNA-Proben werden dann auf einem cDNA-Mikroarray hybridisiert. Dort bindet die mit Farbstoff markierte cDNA an ihren komplementären Gegenpart auf dem Array. Das Fluoreszenzsignale der beiden Farbstoffe von jedem Spot werden mittels eines Laserscanners ausgelesen und in Signalintensitäten umgerechnet.

2.2.2 Bildanalyse

In den vom Mikroarrayscanner aufgenommenen Bildern ist die Information über die Intensitäten der Fluoreszenzfarbstoffe der beiden Kanäle enthalten. Für jeden cDNA-Spot liegen Signalintensitäten sowohl für den Bereich innerhalb des Spots vor, als auch Intensitäten im Hintergrundbereich, die lokal infolge unspezifischer Bindung oder

Verunreinigung des Glasslides entstehen. Zur Analyse der Bilder und Quantifizierung der Signal- und Hintergrundintensitäten werden spezielle Bildanalyseprogramme, zu denen auch das Programm ScanArray gehört, verwendet.
Die Bildanalyse findet in drei Schritten statt. Im ersten Schritt erfolgt zunächst eine Adressierung (*gridding*), die zur Identifizierung der Positionen der Spots führt. Dies geschieht mit einem Raster, welches ein idealisiertes Abbild des Arrays darstellt und sowohl Informationen über die Position jedes Spots als auch Informationen über den jeweiligen geprinteten cDNA-Klon enthält. Die Position des Rasters wird manuell angepasst, da produktionsbedingt leichte Verschiebungen stattfinden. Im nächsten Schritt der Bildanalyse erfolgt die Segmentierung. Hierbei findet die Zuordnung der Pixelintensitäten zu dem Signal bzw. dem Hintergrund der Spots statt (s. Abbildung 2).

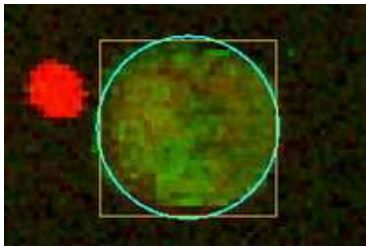


Abb. 2: Darstellung eines Spots mit dem darüber projezierten Bildanalyseraster. Der Kreis stellt den Bereich dar, der als Position für den Spot berechnet wurde. Aus den Pixeln, die sich innerhalb dieses Kreises befinden wird die Signalintensität berechnet. Der rote Punkt ist ein Artefakt im Hintergrund, das sich störend auf die Bildanalyse auswirken könnte.

Im letzten Schritt der Bildanalyse erfolgt die Berechnung der Signal- bzw. der Hintergrundintensitäten aus den Pixelintensitäten der Spots. Neben diesen Werten werden weitere Parameter wie z.B. die Standardabweichung der Pixelintensitäten des Signals berechnet. Diese können als Qualitätskriterium des einzelnen Spots dienen.

Das Programm ScanArray stellt drei verschiedenen Algorithmen für die Bildanalyse zur Verfügung.

Fixed circle

Die *fixed circl*e Methode arbeitet mit konstanten Spotdurchmessern. Entsprechend des verwendeten Rasters werden alle Pixelintensitäten innerhalb der Spotumrandung zur Berechnung der Spotintensität verwendet. Diese Methode eignet sich vor allem, wenn auf den Mikroarrays Spots gleicher Größe an genau definierten Positionen sind.

Adaptive Threshold

Hier erfolgt die Segmentierung durch die Einteilung der Pixelintensitäten in Signal- und Hintergrundpixel in Abhängigkeit eines vom Benutzer gewählten Schwellenwertes.

Histogram

Bei der *Histogram*-Methode wird ein Raster verwendet, dessen Spot-Bereich ein wenig größer ist als die vorhandenen Spots. Die innerhalb des Rasters aufgetragenen Pixelintensitäten werden als Histogramm aufgetragen und aus der Verteilung die Signal- und Hintergrundwerte abgelesen. Diese Methode ist problematisch, wenn ein Raster mit sehr großen Spots, z.B. zum Ausgleichen von produktionsbedingten Unregelmäßigkeiten, zur Auswertung verwendet wird.

2.2.3 Experimentelles Design

Am Anfang eines Experimentes steht die Frage nach dem geeigneten Design. Ein gutes Design soll bei möglichst geringen Kosten eine möglichst hohe Datenqualität erzielen. Das Design muss in Abhängigkeit von der Fragestellung, aber auch im Hinblick auf die Durchführbarkeit und die Auswertung gewählt werden.
Die Varianzen, die in einem Mikroarrayversuch auftreten können, lassen sich in drei Kategorien zusammenfassen.

- biologische Varianzen (z.B. genetische oder Umweltfaktoren, Wachstumsbedingungen)
- technische Variationen oder (z.B. unterschiedliche Hybridisierung, Extraktion Färbung, Qualität des Mikroarrays)
- zufällige Fehler

Alle drei Kategorien können mittels Wiederholungsversuchen kontrolliert werden. Die Anzahl der gewählten Wiederholungs-Mikroarrays ist abhängig von der Streuung der Daten und der zu erwartenden Unterschiede. Systematische Variationen, die z.B. durch unterschiedliches Verhalten der Fluoreszenzfarbstoffe oder Drucknadeln entstehen, können zudem durch eine Normalisierung der Daten (s. Kapitel 2.2.6) korrigiert werden.

Die absolut gemessenen Signalintensitäten eines Zweifarben-Mikroarrayversuches werden durch eine Reihe von Störfaktoren beeinflusst z.B. unterschiedliche Spotgrößen, unterschiedliche Länge der einzelsträngigen DNA Sequenzen oder verschiedene GC-Gehalte der Sequenzen. Von den absoluten Intensitäten der beiden Fluoreszenz-farbstoffe eines Spots kann somit nicht direkt auf die Expressionsstärke der entsprechenden mRNA in den beiden Proben geschlossen werden. Da man annimmt, dass beide Proben im gleichen Maße von den Faktoren beeinflusst werden, wird stattdessen das Intensitätsverhältnis der beiden Proben berechnet.

Bei Zweifarben-Mikroarrays sind deshalb, neben der Anzahl der Wiederholungsversuche, die Kombination der Proben auf den Arrays von Bedeutung. Zwei der am häufigsten verwendeten Designs sind das Paar-Design und das Referenz-Design.

Paar-Design

Beim Paar-Design werden alle zu vergleichenden Proben jeweils gegeneinander hybridisiert. Dieses Design hat aufgrund des direkten Vergleiches der Proben eine geringere Varianz als z.B. das Referenz-Design (s.u.), führt aber bei mehreren Proben zu einer Vielzahl von Gruppierungen und somit zu hohen Kosten. Verwendung findet das Paar-Design häufig beim Vergleich von zwei Proben. Da sich die Intensitätswerte der Fluoreszenzfarbstoffe Cyanin3 und Cyanin5 bei zunehmender cDNA-Menge unterschiedlich verhalten, wird häufig ein Wiederholungsversuch gemacht, bei dem die Farbstoffe getauscht werden, ein sogenannter *dye-swap*-Versuch.

Referenz-Design

Bei dem Referenz-Design werden alle Proben gegen eine allgemeine Referenzprobe hybridisiert. Der Vergleich zwischen zwei Gruppen erfolgt indirekt über das Verhältnis der Proben zur Referenz. Dieses Design wird gewählt, wenn mehrere Gruppen miteinander verglichen werden sollen, aber auch wenn Unterschiede zwischen zwei Gruppen, die mehrere biologische Wiederholungen enthalten, berechnet werden sollen.
Eine Übersicht über das experimentelle Design von Mikroarrayversuchen geben Churchill (2002); Simon and Dobbin (2003) und Yang and Speed (2002).

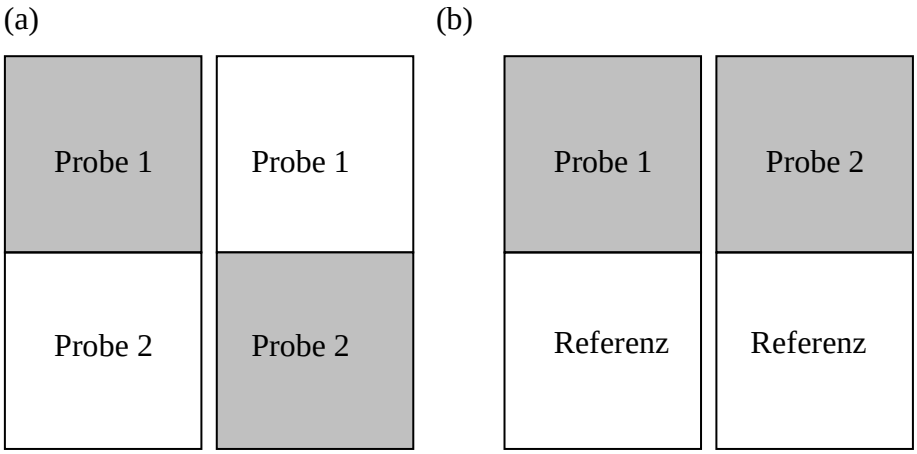


Abb. 3: Schematische Darstellung eines Paar-Designs (a) und des Referenz-Designs (b). (a) Bei dem Paar-Design werden die zu vergleichenden Proben auf dem Mikroarray gegeneinander hybridisiert. Um Farbstoff-abhängige Fehler zu reduzieren, werden in einem Wiederholungsversuch die Farbstoffe getauscht. (b) Bei dem Referenz-Design werden alle Proben gegen eine allgemeine Referenzprobe hybridisiert.

2.2.4 Datentransformation

Die Tatsache, dass nicht die absoluten Intensitäten gemessen werden, sondern die Verhältnisse zwischen den beiden Kanälen, führt dazu, dass hoch- und herunterregulierte Gene in den Zahlenwerten und bei der graphischen Darstellung unterschiedlich behandelt werden. Gene, die um den Faktor 2 hochreguliert werden, haben ein Expressionsverhältnis von 2, wogegen Gene die um den Faktor 2 herunterreguliert werden ein Expressionsverhältnis von 0.5 haben. Die Transformation der Daten auf das duale logarithmische Skalenverhältnis führt zu einer symmetrischen Behandlung der Expressionswerte, so dass ein Gen, welches 2-fach hochreguliert ist, ein $\log_2$-Verhältnis von 1 und ein Gen, welches 2-fach herunterreguliert ist, ein $\log_2$-Verhältnis von -1 hat. Die $\log_2$-Transformation führt somit zur symmetrischen Verteilung der $\log_2$-Verhältnisse um 0. Zusätzlich wird hierdurch die Datenverteilung einer Normalverteilung angepasst. Diese ist eine Grundlage für die Verwendung einer Reihe von statistischen Tests.

2.2.5 Hintergrundkorrektur

Die Intensitäten der einzelnen Spots können durch den lokalen Hintergrund beeinflusst werden. Um diesen Einfluss zu reduzieren, wird die jeweilige lokale Hintergrundintensität von der Signalintensität subtrahiert.

2.2.6 Qualitätskontrolle

Ein wichtiger Schritt der Datenauswertung ist die Qualitätskontrolle und die hieraus resultierende Entscheidung, welche Mikroarrays in die weitere Auswertung und damit in die Beantwortung der gegebenen Fragestellung mit eingehen. Während Mikroarrays mit einer geringen Datenqualität die Streuung der Daten erhöhen können, führt der fälschliche Ausschluss von Mikroarrays zu einer Reduzierung der statistischen Power (Hartmann 2005). Als „statistische Power“ eines Tests bezeichnet man die Wahrscheinlichkeit, mit der ein statistischer Test eine spezifische „richtige“ Alternative unter den Rahmenbedingungen seines Einsatzes (Fallzahl, Signifikanzniveau) auch als solche entdeckt (d.h. die „falsche“ Nullhypothese ablehnt). Zudem ermöglicht die Qualitätskontrolle die Bewertung jedes einzelnen Spots hinsichtlich seiner „Verlässlichkeit“.

Eine Reihe von Parametern werden herangezogen, um die Qualität der Experimente zu beurteilen. Hierbei werden 3 verschiedene Ebenen betrachtet.

- Parameter zur Beurteilung der Qualität der einzelnen Spots
- Parameter zur Beurteilung der einzelnen Arrays
- Parameter zur Beurteilung des gesamten Experimentes

<u>Qualitätsparameter für jeden Spot</u>

Um die Qualität der einzelnen Spots zu bewerten, wird die Spotintensität, das jeweilige Verhältnis der Signal- zur Hintergrundintensität, die Variabilität der Pixelintensität innerhalb eines Spots und die Form des Spots verwendet (Brown, Goodwin et al. 2001; Tseng, Oh et al. 2001; Wang, Klijn et al. 2005).
Spots mit einer geringen Intensität deuten auf eine schwache Hybridisierung der cDNA. Die Signalintensität dieser Spots wird stark vom Hintergrund beeinflusst, was zu einer großen Variabilität der Signalintensitäten im Niedrigintensitätsbereich führt. Solche Datenpunkte sind weniger glaubwürdig als Spots in höheren Intensitätsbereichen. Spots, deren Intensität jedoch nahe der Sättigung liegen, sind ebenfalls wenig glaubwürdig, da die hohen Intensitätswerte häufig durch Verunreinigungen bedingt sind oder aber bei einer Sättigung der Pixelintensitäten durch hohe cDNA-Expressionslevels in diesem Bereich keine realistischen Verhältnisse zwischen den beiden Kanälen mehr gemessen werden können. Eine hohe Variabilität der Pixelintensität innerhalb eines Spots bzw.

eine unregelmäßige Form der Spots weisen ebenfalls auf vorhandene Verunreinigungen hin (Brown, Goodwin et al. 2001).

Eine weitere Möglichkeit, die Qualität der einzelnen Spots zu beurteilen, sind technische Wiederholungen eines Spots mit derselben Probe. Am IMT erfolgt die Hybridisierung bei cDNA Mikroarrays als „Sandwichexperiment". Anstelle eines Deckglases wird hierbei ein zweiter Array verwendet.

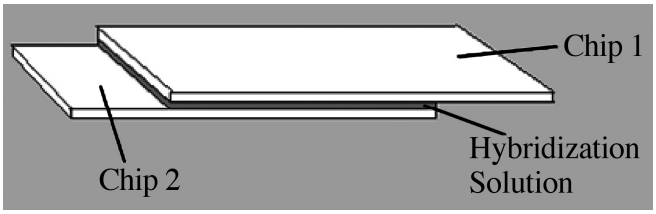


Abb. 4: Schematische Darstellung einer Sandwich-Hybridisierung. Bei der Sandwich-Hybridisierung wird anstelle eines Deckglases ein zweiter Mikroarray verwendet und so eine technische Wiederholungs-messung einer Probe gemacht.

Qualitätsparameter für jeden einzelnen Mikroarray

Eine Reihe von Qualitätsparametern geben Aufschluss über die Qualität der einzelnen Mikroarrays. In der Tabelle 1 sind die Qualitätsparameter beschrieben, die für die Bewertung eines Mikroarrays berechnet werden.

Tab. 1: Qualitätsparameter, die für jeden einzelnen Mikroarray berechnet werden

Fehlende Werte vor/nach Hintergrundkorrektur	Anzahl der Spots, für die vor bzw. nach der Hintergrund-korrektur kein $\log_2$-Verhältnis vorliegt (z.B. wenn in einem der beiden Kanäle kein Intensitätswert vorliegt)
Anzahl der Gene über dem Hintergrund	Anzahl der Spots, welche über dem dritten Quartil der Pixel-intensitäten des lokalen Hintergrundes liegen
Verhältnis Signal/Hintergrund Cy3 und Cy5	Verhältnis der Signalintensitäten zu den Hintergrund-intensitäten aller Spots in den beiden Kanälen
Flag	Anzahl der Spots, die von *ScanArray* mit -100 bewertet wurden. Das Programm bewertet die Qualität jedes Spots mit Werten zwischen -100 (niedrigster Wert) und +100 (höchster Wert). Als Kriterien für die Bewertung kann der Nutzer bei dem Programm *ScanArray* zwischen drei verschiedenen Parametern wählen: • Signal/Hintergrund Verhältnis • Signal/Rauschen Verhältnis • Footprint, d.h. die Übereinstimmung der Position des Spots mit der laut Raster berechneten Position

Eine große Rolle bei der Beurteilung der Arrayqualität spielt die graphische Darstellung der Daten. So geben Bildplots (s. Abbildung 5) verschiedener Parameter, z.B. der Signalintensitäten, der Hintergrundintensitäten oder der $\log_2$-Verhältnisse der beiden Kanäle, Aufschluss über lokale Effekte auf den Chips. Die Bildplots zeigen eine räumliche Darstellung von Intensitätswerten oder $\log_2$-Verhältnisse im xy-Koordinatensystem, bei der die Werte farblich kodiert sind.

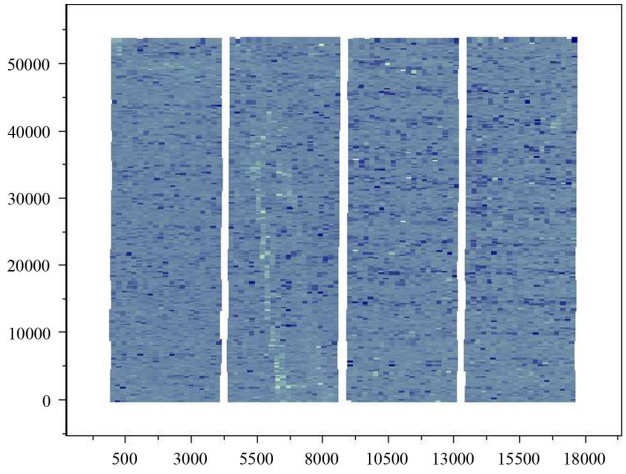


Abb. 5: Bildplot der $\log_2$-Verhältnisse eines Mikroarrays. In dem Bildplot werden die $\log_2$-Verhältnisse der einzelnen Spots entsprechend ihrer Position auf dem Mikroarray farblich kodiert aufgetragen. Je höher der Expressionswert, desto intensiver ist der Blauton. Dies ermöglicht u.a. lokale Unregelmäßigkeiten auf dem Mikroarray zu erkennen. So sind hier im zweiten Block in der Mitte lokale Veränderungen zu sehen, die auf eine Beschädigung des Mikroarray an dieser Stelle hinweisen. (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang)

Mit einem MM-Scatterplot der beiden Sandwich-Mikroarrays wird die Reproduzierbarkeit der Wiederholungsarrays dargestellt. Die $\log_2$-Verhältnisse der beiden Mikroarrays werden hierzu im xy-Koordinatensystem gegeneinander aufgetragen.

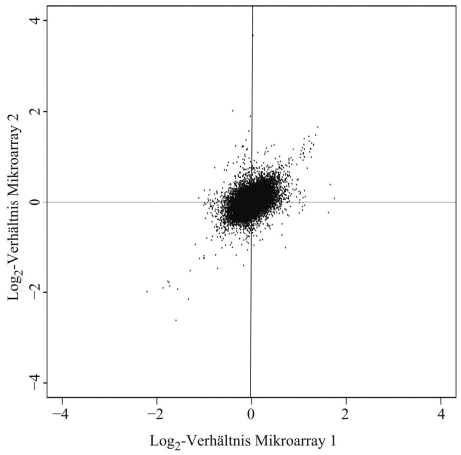


Abb. 6: MM-Plot zweier Mikroarrays. In einem Koordinatensystem werden die $\log_2$-Verhältnisse jedes Spots von zweier Mikroarrays auf der x- und y-Achse gegeneinander aufgetragen. Sie geben Aufschluss über die Reproduzierbarkeit der zwei Mikroarrays. Die Proben der beiden Mikroarrays, die in dieser Abbildung dargestellt sind, zeigen reproduzierbare $\log_2$-Verhältnisse. Dies ist daran zu erkennen, dass die nicht-differentiell exprimierte Gene um den Nullpunkt lokalisiert sind, und die in beiden Mikroarrays reproduzierbar hochregulierten Gene im rechten oberen Bereich liegen, bzw. die herunterregulierten Gene im linken unteren Bereich entlang einer „virtuellen“ Achse, die von links unten nach rechts oben verläuft, liegen. (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang)

Qualitätsparameter zur Bewertung des gesamten Experimentes

Da alle im Experiment verwendeten Proben zumeist einen sehr ähnlichen biologischen Hintergrund haben, sind ähnliche Transkriptionslevel zu erwarten. Mittels verschiedener graphischer Darstellungen aller Mikroarrays eines Experimentes können „Außenseiter"-Arrays erkannt werden. Zu „Außenseiter"-Arrays zählen Mikroarrays mit Proben von geringer Qualität, fehlerhafte Chips oder aber Arrays mit Proben, die einen anderen biologischen/medizinischen Hintergrund haben, da z.B. fälschlicherweise nicht das gewünschte Tumorgewebe, sondern Gewebe aus dem umliegenden Bereich präpariert wurde.

Eine Möglichkeit zur Darstellung des Gesamtexperimentes sind MM-Plots aller Mikroarrays gegeneinander.

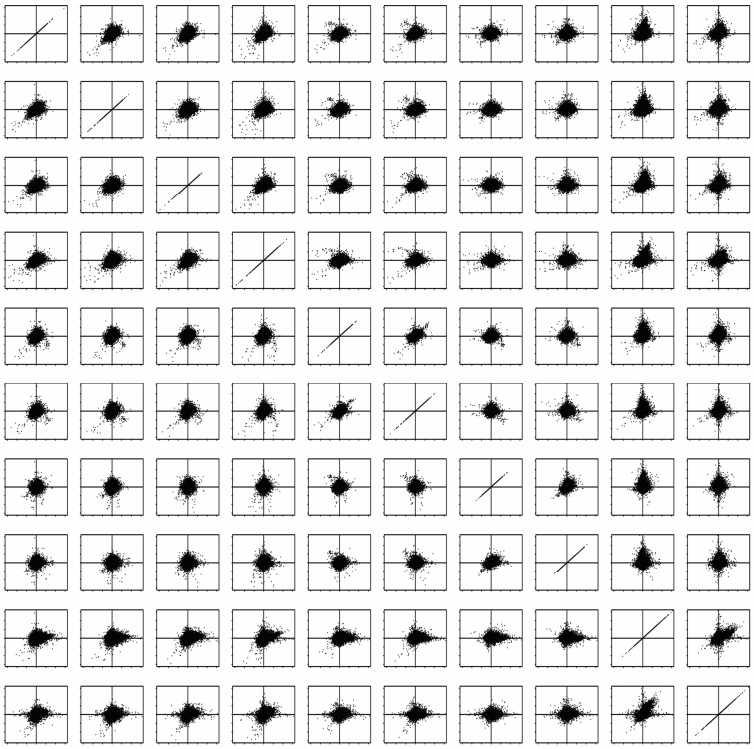


Abb. 7: MM-Plots von 10 Mikroarrays eines Referenz-Design. Die Abbildung zeigt einen MM-Plot von 10 Mikroarrays. Die $\log_2$-Verhältnisse der 10 Mikroarrays werden in einem Koordinatensystem einzeln gegeneinander aufgetragen. Da es sich um Proben mit dem gleichen genetischen Hintergrund handelt, sollten die Expressionswerte der Proben eine gute Reproduzierbarkeit aufweisen d.h. eine „Linie" von links unten nach rechts oben bilden. Die in dieser Abbildung zu beobachtende Streuung der Werte ist neben der technischen Variabilität vor allem auf die biologische Variabilität der unterschiedlichen Mäuse zurückzuführen. (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang. Sie zeigt die 10 Mikroarrays auf denen die Tumore hybridisiert sind, welche das Myc-Wildtyp Allel tragen).

Eine weitere Methode zum Identifizieren von „Außenseiter"-Arrays innerhalb eines Experimentes ist die Darstellung der Daten als Boxplots oder das Cluster-Verfahren.

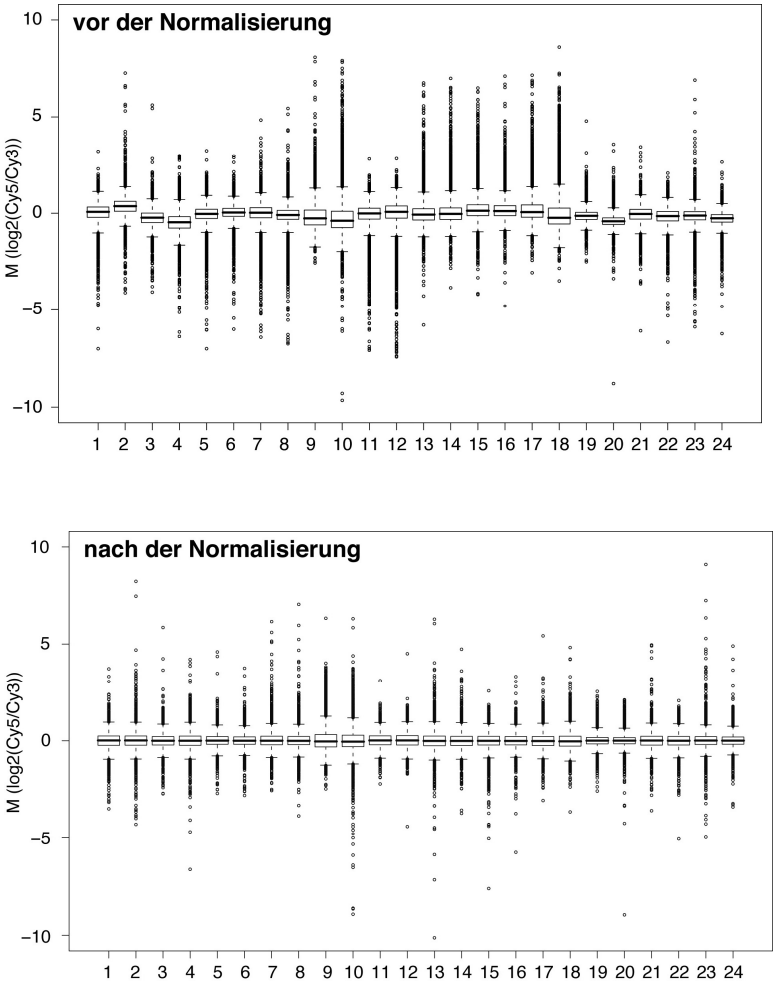


Abb. 8: Boxplots der log_2 Verhältnisse eines Versuches vor und nach der Normalisierung. Mit Boxplots wird die Häufigkeitsverteilung der Daten dargestellt. Die untere und obere Linie der einzelnen Boxplots stellen das erste und dritte Quartil dar, die mittlere Linie den Median der log_2-Verhältnisse des jeweiligen Mikroarrays dar. Das Ende der Linien über und unterhalb der Box zeigen das Maximum beziehungsweise das Minimum einer Verteilung an. Werte, die mehr als das 1,5-fache des Interquartilabstands vom Median abweichen, werden als Ausreißer bewertet und werden als einzelne Datenpunkte dargestellt. Da die Proben den gleichen genetischen Hintergrund haben, sollten sie ähnliche Verteilungen aufweisen d.h. die Boxplots sollten vor allem einen ähnlichen Quartilsabstand aufweisen. Der Bereich der log_2-Verhältnisse wird u.a. durch systematische Faktoren beeinflusst, die mittels der Normalisierung korrigiert werden können (s. 2.2.6). Die dargestellten Boxplots zeigen alle einen ähnlichen Quartilsabstand. Die Boxplots liegen vor der Normalisierung in unterschiedlichen Bereichen, so sind z.B. Mikroarray 4 und 5 niedriger als die Mikroarrays 1 und 2. Infolge der Normalisierung sind diese Unterschiede angeglichen worden. (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang)

Boxplots stellen die Verteilung der log_2-Verhältnisse der gesamten Werte jedes Chips dar. Diese sollten – nach der Normalisierung – eine ähnliche Verteilung bzw. ein

ähnliches Intensitätslevel aufweisen. In Abbildung 8 sind die Boxplots eines Versuches mit 24 Mikroarrays vor und nach der Normalisierung dargestellt.

Das Clusterverfahren ist ein unüberwachtes Analyseverfahren zur Ermittlung von Gruppen (Clustern) von Versuchsproben, deren Eigenschaften oder Ausprägungen bestimmte Ähnlichkeiten oder Unterschiede aufweisen. Das Verfahren wird im Detail in Kapitel 2.2.9 beschrieben. Es ist davon auszugehen, dass Mikroarrays, auf denen technische bzw. biologische Wiederholungsproben hybridisiert sind, ein gemeinsames Cluster bilden. Somit sollten die im Versuch betrachteten unterschiedlichen Gruppen auch getrennte Cluster bilden. Bei Mikroarrays, die in eine andere Gruppe geclustert werden, könnte es sich dann um „Außenseiter"-Arrays handeln (Brown, Goodwin et al. 2001; Tseng, Oh et al. 2001; Wang, Ghosh et al. 2001).

2.2.7 Normalisierung

Verschiedene Faktoren können innerhalb von Mikroarray-Versuchen zu systematischen Abweichungen führen. Hierzu gehören:

- Unterschiedliche Mengen an eingesetzter RNA/cDNA in den beiden Kanälen
- Sättigung (Scanner; Labelling)
- Nicht-Linearität der Cyanin5- und Cyanin3-Färbung
- Effizienzen der Cyanin5- und Cyanin3-Färbung
- Variation im Niedrigintensitätsbereich
- Printernadeln
- Variierende PCR-Reaktionen der verschiedenen Platten
- Lokale Effekte auf den Arrays, z.B. Verunreinigungen

Diese systematischen Abweichungen erschweren einerseits die Unterscheidung von regulierten und nicht-regulierten Genen, andererseits die Vergleichbarkeit der Resultate verschiedener Arrays. Methoden, die solche systematischen Abweichungen korrigieren, bezeichnet man als Normalisierungsmethoden. Ziel ist es, die Auswirkungen der systematischen Abweichungen zu minimieren, um so besser auf die biologischen Veränderungen fokussieren zu können und die Vergleichbarkeit der verschiedenen Arrays zu erhöhen.

Es gibt verschiedene Methoden zur Normalisierung von Mikroarrayexperimenten, die generell in zwei Gruppen eingeteilt werden können.

Normalisierung basierend auf der Verteilung des Gesamtdatensatzes

- Globale Normalisierung
- Lowess Normalisierung
- Print-tip-Lowess Normalisierung

Normalisierung basierend auf einer Teilmenge der Gene

- Housekeeping Gene
- Spike-In-Kontrollen

Normalisierung basierend auf der Verteilung des Gesamtdatensatzes

Der Normalisierung über die Verteilung des Gesamtdatensatzes liegt die Annahme zugrunde, dass nur ein geringer Teil der Gene innerhalb des Mikroarrayversuches differentiell exprimiert sind oder dass die Expressionsraten hoch- und herunterregulierter Gene symmetrisch verteilt sind. Weiterhin sollten eine möglichst große Anzahl von Datenpunkten für jeden Mikroarray vorliegen, d.h. die Anzahl der Spots auf dem Mikroarray ausreichend groß sein, um systematische Fehler ausreichend darstellen zu können.

Globale Normalisierung

Die globale Normalisierung basiert auf folgenden Annahmen:
In jeder Probe wird die gleiche Menge an RNA mit der gleichen Anzahl an RNA Molekülen eingesetzt. Die verwendeten Proben repräsentieren eine Zufallsauswahl von Genen eines Organismus, so dass die meisten Gene unverändert sind. Hieraus resultiert die Annahme, dass beide Kanäle eines Mikroarray-Objektträgers durch einen konstanten Faktor miteinander verknüpft sind. Dieser Faktor kann berechnet werden, indem man die Summe der gemessenen Intensitäten der beiden Kanäle berechnet und ins Verhältnis setzt. Zur Darstellung der systematischen zu korrigierenden Fehler und den Veränderungen, die durch das Normalisierungsverfahren erzielt werden, eignet sich der MA-Plot. Dieser stellt auf der x-Achse den A (*average*) -Wert dar, welcher der Mittelwert der absoluten Intensität der beiden Kanäle für jeden Spot ist.

$A_i=0.5*(\log_2 Cy5+\log_2 Cy3)$

Cy5= Intensität des Spot im Cyanin5-Kanal

Cy3= Intensität des Spot im Cyanin3-Kanal

und auf der y-Achse den M-Wert, der das Verhältnis der beiden Kanäle darstellt.

$M_i=\log_2 Cy5-\log_2 Cy3$

Cy5= Intensität des Spot im Cyanin5-Kanal

Cy3= Intensität des Spot im Cyanin3-Kanal

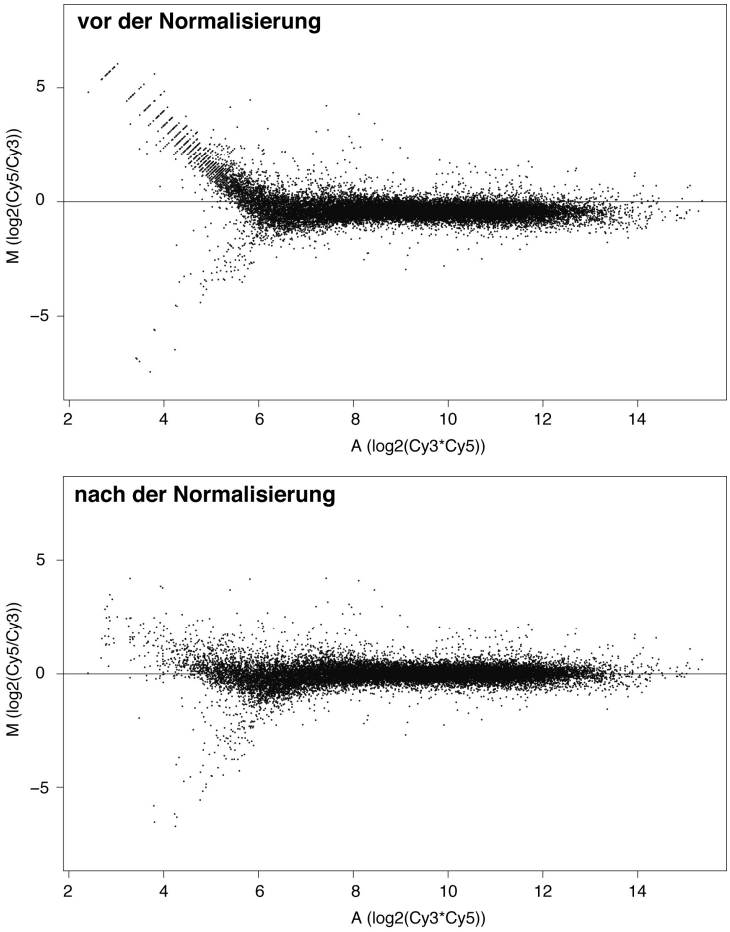


Abb. 9: MA-Plots eines Datensatzes vor und nach der globalen Normalisierung. Der MA-Plot stellt auf der x-Achse die mittlere $\log_2$-Intensität der beiden Kanäle eines jeden Spots als A-Wert und auf der y-Achse das Verhältnis der $\log_2$ Intensitäten der beiden Kanäle als M-Wert dar. Vor der Normalisierung liegen die M-Werte vorwiegend oberhalb der x-Achse. Durch die Normalisierung erfolgt eine Verschiebung der $\log_2$-Verhältnisse „nach unten“, so dass sie symmetrisch um die x-Achse streuen. Im Niedrigintensitätsbereich liegt häufig eine höhere Streuung der Werte vor, da sie stärker vom Hintergrund beeinflusst werden. (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang).

Abb.9 zeigt denselben MA-Plot vor (a) und nach (b) der globalen Normalisierung. In Abbildung (a) sieht man, dass die Regressionsgerade der M-Werte oberhalb der x-Achse liegt und nach der Normalisierung symmetrisch um die x-Achse streut. Die Normalisierung der gemessenen $\log_2$-Verhältnisse mit diesem Korrekturfaktor führt dazu, dass das mittlere $\log_2$-Verhältnis über alle Gene 0 ist.
Eine globale Normalisierung ist jedoch nur sinnvoll, wenn beide Kanäle durch einen konstanten Faktor gekoppelt sind. Dies ist häufig nicht der Fall da intensitätsabhängige Effekte auftreten.

Intensitätsabhängige Normalisierung

Lowess-Normalisierung

Intensitätsabhängige Effekte haben z.B. ihr Ursache in den unterschiedlichen physikalischen Eigenschaften der Cyanin-3- und Cyanin-5-Farbstoffe (Stabilität bei höheren Ozonwerten und Temperaturen, unterschiedliche Effizienzen in der Farbstoffeinlagerung). Wenn solche Effekte vorliegen, sollte eine intensitätsabhängige Lowess-(*locally weighted linear regression analysis*) Normalisierung durchgeführt werden (Yang, Dudoit et al. 2002). Die Lowess-Normalisierung ist eine nicht-parametrische Prozedur, welche für das „Glätten“ von Scatterplots verwendet wird. Bei dem Zweifarben-Mikroarray wird es zum „Glätten“ des MA-Plots verwendet (s. Abbildung 10). Eine Lowess-Normalisierung sollte nur durchgeführt werden, wenn eine ausreichende Anzahl von Genen auf dem Mikroarray sind, die eine Abschätzung intensitätsabhängiger Regressionsgeraden erlauben.

Print-Tip-Lowess-Normalisierung

Eine häufige Ursache für systematische Abweichungen ist das unterschiedliche Printergebnis der verschiedenen Nadeln bei der Herstellung der cDNA-Mikroarrays. Die Korrektur erfolgt hierfür, indem für jede Gruppe von Spots, die mit der gleichen Nadel hergestellt wurden, intensitätsabhängige Regressionsgeraden erstellt werden, über die dann normalisiert wird.

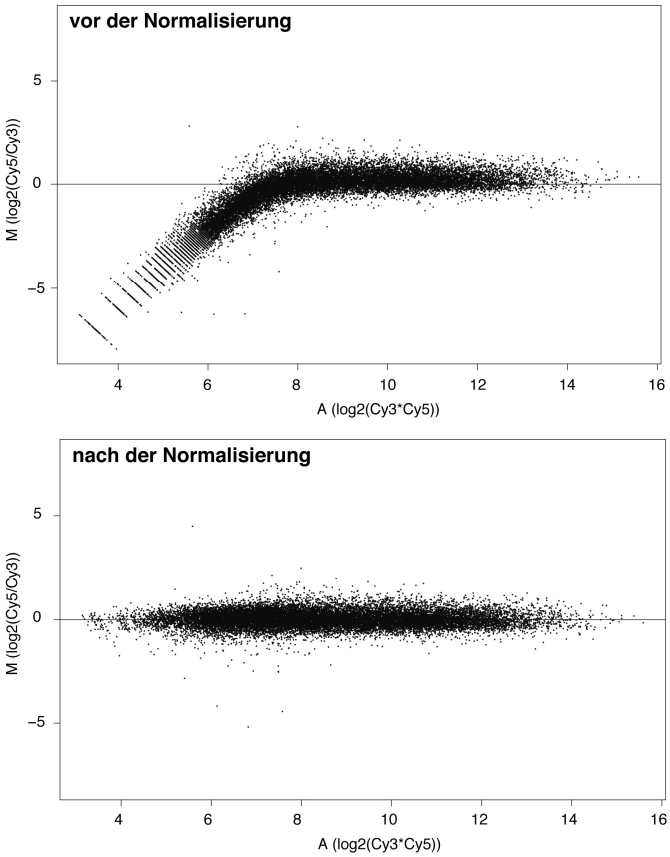


Abb. 10: MA-Plots eines Datensatzes vor und nach der Lowess-Normalisierung Auf dem MA-Plot wird auf der x-Achse der A-Wert d.h. die mittlere $\log_2$-Intensität der beiden Kanäle eines jeden Spots und auf der y-Achse der M-Wert d.h. das Verhältnis der $\log_2$ Intensitäten der beiden Kanäle dargestellt. Vor der Normalisierung zeigt die „Punktwolke" im niedrigen Intensitätsbereich eine Verschiebung zu niedrigeren $\log_2$-Verhältnissen. Diese intensitätsabhängigen Effekte werden durch die Lowess-Normalisierung aufgehoben, die „Punktwolke" liegt symetrisch um die x-Achse (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang).

Normalisierung über ein Subset von Genen

Wenn die Annahmen, die für die Normalisierung über die Verteilung des Gesamtdatensatzes gelten, nicht vorliegen oder es sich um einen „kleinen" Mikroarray mit nur wenigen Spots handelt, so dass eine Verteilung nicht abgeschätzt werden kann, sollte über Teilmengen der Gene normalisiert werden. Hierbei wird angenommen, dass dieses Subset an Kontrollgenen in beiden auf dem Array hybridisierten Proben gleich exprimiert wird oder aber im gleichen Maße zu beiden Proben hinzupipettiert wurde. So kann das Verhältnis dieser Gene als Basis für die Normalisierung des Gesamtdatensatzes verwendet werden. Nach der Normalisierung sind die Expressionsverhältnisse der Kontrollgene um Null angeordnet. Als Subsets wurden bisher *Housekeeping*-Gene und *Spike-In*-Kontrollen verwendet.

Housekeeping-Gene

Housekeeping-Gene sind Gene, von denen man weiß bzw. annimmt, dass sie unter experimentellen Bedingungen konstant exprimiert werden. Da sie aber häufig dennoch differentiell exprimiert werden und zudem nicht den gesamten Intensitätsbereich des Experimentes abdecken, sollten sie nur in Ausnahmefällen zur Normalisierung verwendet werden.

Spike-In-Kontrollen

Spike-In-Kontrollen sind cDNA-Sequenzen, die nicht in den Proben vorhanden sind und in genau definierten gleichen Mengen jeder einzelnen Probe hinzu gegeben werden. Die Hinzugabe von Kontrollen zu verschiedenen Zeitpunkten des Experimentes fungiert zusätzlich als positive Kontrolle zur Überprüfung der einzelnen experimentellen Schritte. Da die Sequenzen in gleichen Mengen in beiden Proben vorliegen, sollte ihr Verhältnis bei Null liegen und damit als Basis zur Normalisierung dienen. Wichtig ist auch hier, dass die Kontrollen den gesamten Intensitätsbereich abdecken.

2.2.8 Selektion differentiell exprimierter Gene

Die Herausforderung bei der statistischen Analyse von Mikroarray-Experimenten besteht in der hohen Anzahl an getesteten Genen (Merkmalen) im Vergleich zu der häufig nur geringen Anzahl an untersuchten Proben (Beobachtungseinheiten).
Die Entscheidung, ob ein Gen differentiell exprimiert ist, beinhaltet zwei Schritte. Zunächst ist eine Statistik auszuwählen, um die Gene anhand der Unterschiede zwischen den verschiedenen Phänotypen zu gewichten. Im zweiten Schritt muss ein Kriterium gewählt werden, anhand dessen entschieden wird, welche der Gene als differentiell exprimiert ausgewählt werden.

Fold change

Im einfachsten Fall fungiert die relative Änderung (*Fold change)* der Expressionswerte zwischen den zu vergleichenden Proben, die dem delogarithmierten M-Wert entspricht, als Rankingkriterium. Dieser Ansatz berücksichtigt jedoch nur die Unterschiede und vernachlässigt die Varianz innerhalb der einzelnen Phänotypen. Ein Ranking anhand des *Fold changes* wird deshalb hauptsächlich bei sehr kleinen Gruppengrößen gewählt.

Als Kriterium, welche Gene differentiell exprimiert sind, wird der *Fold change* gewählt, bei dem die Gene im MA-Plot außerhalb der Gesamtstreuung liegen (s. Abbildung11).

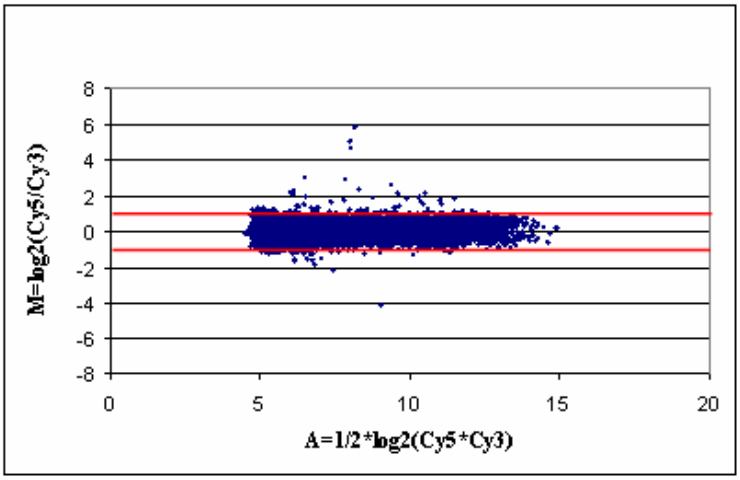


Abb. 11: MA-Plot zur Selektion von differentiell exprimierten Genen anhand des *Fold changes*. Der MA-Plot stellt auf der x-Achse den A-Wert und auf der y-Achse den M-Wert eines jeden Spots dar. Gene, die über bzw. unterhalb der roten Linie sind, haben eine $\log_2$-Verhältnis, das größer als 1 bzw. kleiner als -1 ist und werden als differentiell exprimiert selektiert (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang).

Volcanoplot

Um die Varianz auch bei kleinen Fallzahlen zu berücksichtigen, verwendet man den *Volcanoplot*. Hier wird als Selektionskriterium für jedes Gen, neben dem log2-Verhältnis, auch der t-Wert einer Teststatistik verwendet. Teststatistiken sind Hypothesentests, die überprüfen, ob ein gefundener Mittelwertunterschied rein zufällig entstanden ist, oder ob es wirklich bedeutsame Unterschiede zwischen den zwei untersuchten Gruppen gibt. Hierbei wird neben den Mittelwertsunterschieden die Streuung der Werte innerhalb der Gruppen berücksichtigt. So erfolgt z.B. die Berechnung der t-Statistik für zwei unabhängige Stichproben mit der folgenden Formel:

$$t_{n_1+n_2-2} = \frac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}_{\overline{x}_1-\overline{x}_2}}$$

$\overline{x}_1 - \overline{x}_2$ = Differenz der Mittelwerte der beiden Gruppen

$\hat{\sigma}_{\overline{x}_1-\overline{x}_2}$ = Differenz der Standardabweichungen der beiden Gruppen

In einem Scatterplot wird für jeden cDNA-Klon der M-Wert auf der x-Achse gegen den absoluten t-Wert auf der y-Achse aufgetragen. Gene, die sowohl einen großen *fold change* als auch einen hohen t-Wert aufweisen, werden als differentiell exprimiert selektiert.

Anhand des *Volcanoplots* kann ähnlich wie beim MA-Plot die Streuung der Daten abgeschätzt werden und in die Entscheidung, welche der Gene selektiert werden, mit einbezogen werden.

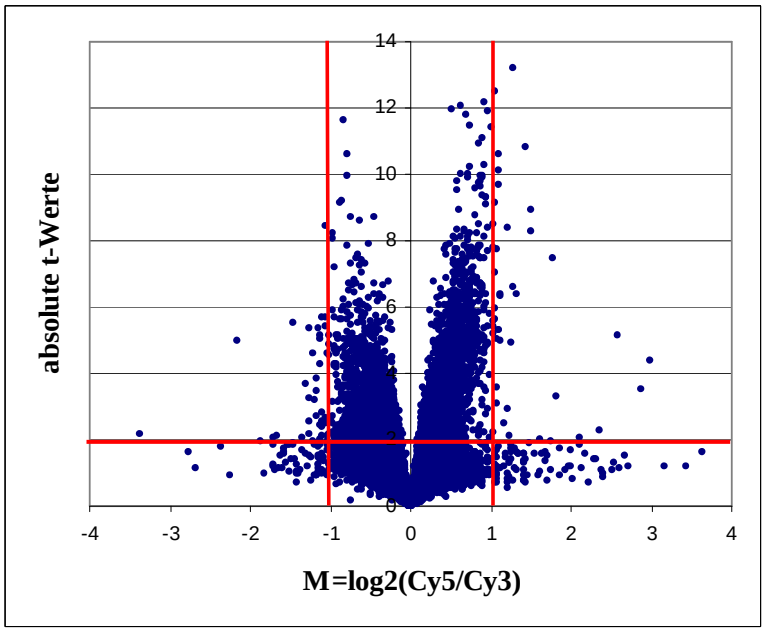


Abb. 12: *Volcanoplot* zur Selektion von differentiell exprimierten Genen. Bei dem *Volcanoplot* werden die $\log_2$-Verhältnisse auf der x-Achse gegen die absoluten t-Werte auf der y-Achse aufgetragen. Als differentiell exprimiert werden Gene selektiert, die sowohl einen großen *Fold change* als auch einen hohen t-Wert aufweisen. Die Selektionsgrenzen für beide Parameter werden durch die roten Linien dargestellt (Abbildung stammt aus Experiment 1, siehe Experimentenliste im Anhang).

Da beide Methoden mit der Festlegung der Schwellenwerte durch den Anwender nur ein willkürlich gewähltes Maß für die Selektion differentiell exprimierter Gene ermöglichen, sollten sie nur bei einer kleinen Fallzahl verwendet werden.

t-Statistik

Bei größeren Fallzahlen kann für jeden cDNA-Klon des Mikroarrays eine t-Statistik zum Vergleich der Expressionswerte der verschiedenen Gruppen durchgeführt werden. Diese berücksichtigt neben den Gruppenunterschieden auch die Varianz innerhalb der Gruppen. Problematisch bei der Verwendung der t-Statistik kann jedoch sein, dass bei

cDNA-Klonen, die innerhalb der verschiedenen Gruppen nur eine sehr kleine Standardabweichung haben, sich hohe t-Werte ergeben, selbst wenn die Gruppenunterschiede nur klein sind. Aufgrund der Vielzahl der durchgeführten Tests ist die Wahrscheinlichkeit, dass dies auftritt, recht groß. Weiterhin wird bei der Ableitung der p-Werte aus der t-Verteilung eine Normalverteilung vorausgesetzt, was häufig aufgrund der geringen Gruppengrößen nicht zutrifft.

SAM (*significance analysis of microarrays*)

Die SAM-Methode (Tusher, Tibshirani et al. 2001) versucht dieses Problem zu umgehen, indem sie bei der Berechnung der Statistik anstelle der realen Standardabweichung einen Wert verwendet, bei dem zur Standardabweichung im Nenner der Gleichung eine kleine Konstante hinzuaddiert wird, die aus der Standardabweichung der Werte über den ganzen Mikroarray abgeleitet wird. Die p-Werte werden über ein Permutationsverfahren berechnet, das keine Normalverteilung der Daten voraussetzt. Die Korrektur für das multiple Testproblem (s.u.) erfolgt durch die Berechnung der *false discovery rate*.

2.2.9 Multiples Testproblem

Je mehr Gene parallel getestet werden, desto höher ist die Wahrscheinlichkeit, dass trotz der Gültigkeit der Nullhypothese für jeden einzelnen Test, die besagt, dass die Wahrscheinlichkeit, dass ein Gen falsch positiv als differentiell exprimiert selektiert wird, unterhalb des Signifikanzniveaus liegt, der Anteil der falsch positiven Gene deutlich höher liegt. Um nicht zu falsch-positiven Resultaten zu kommen, muss in diesem Fall eine entsprechende Korrektur des Signifikanzniveaus im Einzeltest durchgeführt werden. Es gibt verschiedene Ansätze das multiple Testproblem zu kontrollieren. Eine Übersicht hierzu geben Bender, Lange et al. (2007), Efron et al. (2001) und Storey et al. (2001).

2.2.10 Clustering und Visualisierung

Die Clusterverfahren gehören zu den explorativen unüberwachten Analyseverfahren und dienen zum Auffinden von bisher nicht bekannten Zusammenhängen. Diese Verfahren spielen bei der Auswertung von Mikroarrayversuchen eine wichtige Rolle,

z.B. zum Auffinden von neuen Subtypen von Krankheiten. Sie fungieren primär zur Datenreduktion bzw. Visualisierung, spielen aber auch bei der Qualitätskontrolle und der Etablierung neuer Mikroarrays-Plattformen eine wichtige Rolle.

Ziel ist es, die Gruppierungen zwischen Genen oder Experimenten auf Basis von Ähnlichkeitsmerkmalen zu finden. Die Ähnlichkeit zwischen den Gruppen soll hierbei möglichst groß und die Ähnlichkeit innerhalb der Gruppen möglichst gering sein. Generell kann man die verschiedenen Clusterverfahren in hierarchische und partitionierende Verfahren unterteilen. Es gibt eine Reihe von Methoden für das hierarchische Clustern, die sich in der Ähnlichkeitsmetrik (z.B. euklidische Distanz, Manhattan-Distanz), d.h. der Abstandberechnung zwischen den einzelnen Genen bzw. den Experimenten, und den Fusionsverfahren, d.h. dem Abstand zwischen den einzelnen entstehenden Clustern (*single-, complete-, average-distance*) unterscheiden.

Bei den partitionierenden Clustern wird die Anzahl der Cluster (k) vom Anwender vorgegeben. Die Aufteilung beginnt dann mit k Datenpunkten, welche als sogenannte ´centroids` fungieren. Der Algorithmus teilt dann die Objekte in Cluster, indem er sie schrittweise dem nächsten Centroid-Cluster zuteilt. Nachdem alle Objekte auf die Centroids aufgeteilt wurden, werden die Centroids-Mittelpunkte berechnet. Anschließend wird jedes Objekt wieder seinem nächsten Cluster zugeordnet. Dies wird wiederholt, bis keine Veränderung mehr auftritt. Die k-*means* Clustermethode und SOM (*self organizing maps*) gehören zu dieser Gruppe. Einen Überblick über Clustermethoden findet man bei *Algorithms for clustering data* von A.K. Jan.

3 Funktionelle Analyse eines Mikroarray-Datensatzes

3.1 Einleitung

Die Mikroarray-Technologie gibt es seit Beginn der 90er Jahre. Die ersten Ergebnisse von cDNA-Mikroarray-Experimenten wurden 1995 publiziert (Schena, Shalon et al. 1995), 1996 folgten dann die ersten Ergebnissen von Affymetrix GeneChips® (Lockhart, Dong et al. 1996). Inzwischen sind Genexpressionsstudien ein fester Bestandteil der Genomforschung. Während die Erstellung von Genexpressionsprofilen in vielen Labors zur Routine gehört, gestaltet sich die Interpretation der Daten jedoch immer noch als sehr schwierig. Bei der Auswertung von Ergebnissen wird häufig zunächst auf die Gene, die am stärksten hoch- oder herunterreguliert sind, bzw. auf Gene mit einem interessanten biologischen Hintergrund fokussiert. Dieser Ansatz ist jedoch nicht optimal, da bei der Betrachtung von nur einzelnen Genen Informationen verloren gehen, wie z.B. die Zugehörigkeit eines hohen Anteiles der Gene der Ergebnisliste zu bestimmten medizinischen oder biologischen Themenbereichen. Eine weitere Problematik bei der Interpretation von Daten ist häufig, verschiedene Mikroarray-Versuche zur gleichen biologischen Fragestellung miteinander in Bezug zu setzen, da die Ergebnislisten keine oder kaum Übereinstimmungen zeigen.
Um die Interpretation der Ergebnisse zu verbessern, wurde in den letzten Jahren eine Reihe von bioinformatischen Methoden etabliert. Diese ermöglichen es, die Gene der Ergebnislisten bestimmten funktionellen Bereichen zuzuordnen, bzw. sie mit den Ergebnissen anderer Genexpressionstudien oder Genlisten, die anhand biologischer Gemeinsamkeiten zusammengestellt wurden, z.B. Gene eines Stoffwechselweges, zu vergleichen.

Bei der Interpretation von Ergebnislisten auf funktionelle Gemeinsamkeiten ist zu beachten, dass die Ko-Regulation von Genen auf verschiedene Ursachen zurückgeführt werden kann.

So werden Gene ko-reguliert, wenn sie

- durch einen gemeinsamen Transkriptionsfaktor reguliert werden
- in einer stromaufwärts- oder stromabwärts-Beziehung stehen, in der die Produkte der stromaufwärts liegenden Gene die Aktivität der stromabwärts liegenden Gene regulieren
- in einen gemeinsamen Stoffwechselweg involviert sind

In diesem Kapitel werden folgende Methoden zur funktionellen Analyse von Genexpressionsdaten beschrieben.

- *Expression Analysis Systematic Explorer* (EASE)
- *Gene Set Enrichment Analysis* (GSEA)
- *Connectivity Map*
- *Ingenuity Pathway Analysis*
- Untersuchung der Promotorsequenzen der Gene auf gemeinsame cis-regulatorische Motive

3.2 Methoden zur Untersuchung von Genlisten auf funktionelle Gemeinsamkeiten

3.2.1 Gene Ontology Terms

Eine wichtige Aufgabe bei der Interpretation von Genexpressionsprofilen ist eine möglichst umfassende Annotierung der Ergebnislisten. Hierzu gehört neben der Zuordnung entsprechender Datenbank-Identifikationsnummern (*Genbank ID, Unigene ID, GeneID*) die funktionelle Annotierung der Gene. Dies erfolgt meistens auf Basis der *Gene Ontology terms*.

Das *Gene Ontology* Projekt

Die *Gene Ontology* Klassifizierungen werden durch das Gene Ontology Konsortium festgelegt. Dies ist eine Gruppe von Wissenschaftlern, die Gene und Genprodukte mittels eines bestimmten festgelegten Vokabulars – den so genannten *Gene ontology terms* – beschreiben und so die Standardisierung der Beschreibung der Funktionen von Genen und Genprodukten erreichen (Ashburner, Ball et al. 2000; 2001; 2006).

Es werden hierbei drei Kategorien unterschieden:

- Molekulare Funktion der Gene (*molecular function*)
- Biologische Prozesse (*biological process*) – die Rolle der Gene in biologischen Prozessen
- Zelluläre Komponenten (*cellular component*) – die zelluläre Lokalisation und die Beteiligung an zellulären Strukturen

Ein Genprodukt kann eine oder mehrere Funktionen haben, an einem oder mehreren biologischen Prozessen teilnehmen und in einem oder mehreren zellulären Kompartimenten vorkommen.

3.2.2 DAVID/EASE

Die DAVID (*Database for Annotation, Visualization and Integrated Discovery*) Software wurde vom *National Institute of Allergy and Infectious Diseases* (NIAID) etabliert und steht auf einer web-basierten Plattform zur Verfügung. Zur lokalen Anwendung der Funktionen kann die EASE (*Expression Analysis Systematic Explorer*) Software von dem Server des NIAID (http://david.abcc.ncifcrf.gov/) heruntergeladen und installiert werden. Mittels dieser Software können vorherrschende biologische Themen einer Genliste detektiert werden.

Die derzeitige Version stellt folgende Kategorien zur Verfügung:

- Stoffwechselwege aus den Datenbanken
  - *Kyoto Encyclopedia of Genes and Genomes* (KEGG)
  - *Biological Biochemical Image Database* (BBID)
  - *Gene Map Annotator and pathway profiler* (GENMAPP)
- Gene Ontology terms
- Proteinfunktionen aus den Datenbanken
  - *Protein Information Resource* (PIR)
  - SWISSPROT
- Chromosomale Lokalisation
- Enzymatische Funktionen basierend auf der EC Nummer
- Proteindomänen aus den Datenbanken
  - Interpro

- PFAM
- PIR
- SMART

Mit dem EASE *score* – ein Signifikanzwert, der auf einem modifizierten Fisher-Exakt-Test beruht, wird für jede Kategorie angegeben, ob die Gene der Ergebnisliste, die diesem Begriff zugeordnet werden, in Bezug zu allen im Datenset vorhandenen Genen signifikant häufiger vertreten sind.

3.2.3 Gene Set Enrichment Analysis (GSEA)

Die *Gene Set Enrichment Analysis* (GSEA) ist eine Methode für die Untersuchung von im Vorfeld festgelegten Gengruppen auf signifikante Unterschiede zwischen verschiedenen Phänotypen (Subramanian, Tamayo et al. 2005). Bei den Gengruppen handelt es sich zum Beispiel um

- Gene, die dem gleichen Stoffwechselweg oder Signalweg angehören
- Gene eines *Gene Ontology-Terms*
- Gene, die auf dem gleichen Chromosom liegen
- Gene, die ein bestimmtes regulatorisches Motiv im Promotorbereich/3´-UTR-Bereich haben
- Ergebnislisten von anderen Genexpressionsstudien
- Genlisten, die Gene beinhalten, deren Expression mit bekannten Onkogenen korreliert. Basierend auf einer Liste von bekannten Tumor-assoziierten Genen wurden Genexpressionsdatenbanken nach Genen durchsucht, die mit diesen Onkogenen korrelieren. Es wurden die Genexpressionsprofile der folgenden drei Kompendien verwendet:
  - MORF *National Cancer Institute* (http://dtp.nci.nih.gov)
  - GNF2 *Novartis Research Foundation SymAtlas* (Su, Wiltshire et al. 2004)
  - GCM *Global Cancer Map* (Ramaswamy, Tamayo et al. 2001).

Bei der GSEA-Methode wird das gesamte Genexpressionsprofil verwendet. Dies hat den Vorteil, dass Gene, die nur einen geringen Unterschied zwischen den Phänotypen aufweisen, aber möglicherweise in Ko-Regulation mit anderen Genen agieren, nicht schon im Vorfeld von der Analyse ausgeschlossen werden.

Zunächst werden die Gene entsprechend dem Rankingkriterium, das bei der Auswertung verwendet wurde, sortiert. Gensets, die von der Software vorgegeben werden, aber auch selbst erstellt werden können, werden mit der sortierten Liste verglichen, um zu überprüfen, in welchem Maße die Gene der einzelnen Sets am oberen oder unteren Ende dieser Liste angereichert sind. Die Berechnung des sogenannten *enrichment scores* erfolgt über eine Rangsummenstatistik. Hierbei wird an der Liste absteigend entlang gegangen und der Wert der Rankingstatistik hinzuaddiert, wenn das Gen im Genset enthalten ist und subtrahiert, wenn es nicht in dem Genset vorhanden ist. Der *enrichment score* ist die maximale Abweichung von Null, die im Laufe des Vorganges berechnet wird und anschließend auf die Anzahl der Gene in einem Genset korrigiert wird. Das Signifikanzlevel des *enrichment scores* wird durch ein Permutationsverfahren berechnet, welches auf dem Phänotyp basiert. Eine mögliche Korrelation der Gene wird hierbei berücksichtigt. Die Korrektur des multiplen Testproblems erfolgt über die Berechnung der *false discovery rate* (s. Kapitel 2.2.8).

3.2.4 Connectivity Map

Die *Connectivity Map* (http://www.broad.mit.edu/cmap/) stellt molekulare Zusammenhänge zwischen den Ergebnissen von Genexpressionsversuchen und der Wirkung kleiner bioaktiver Moleküle auf Zelllinien her. Die Software basiert auf einer Datenbank, in der Genexpressionsprofile von humanen Zelllinien, die mit verschiedenen bioaktiven Molekülen behandelt wurden, gespeichert sind. Die derzeitige Version von *Connectivity Map* beinhaltet Genexpressionsstudien von 164 verschiedenen kleinen Molekülen, die in verschiedenen Konzentrationen getestet wurden, so dass insgesamt 453 Profile vorliegen. Der Anwender gibt die Ergebnisliste des Mikroarrayversuches getrennt nach hoch- und herunterregulierten Genen ein. Die Software untersucht dann die Verteilung dieser Gene in jedem einzelnen Genexpressionsprofil der Datenbank (Lamb 2007). Die Berechnung einer möglichen Anreicherung erfolgt entsprechend der im vorherigen Kapitel beschriebenen GSEA-Methode.

3.2.5 Ingenuity Systems

Ingenuity Pathway Analysis (http://www.ingenuity.com/) ist eine web-basierte Software, die aus Genlisten mittels eines dynamischen Algorithmuses mögliche

Netzwerke dieser Gene berechnet. Für jedes Gen wird innerhalb der Ergebnisliste nach beschriebenen Interaktionpartnern gesucht. Die erfolgt auf der Basis einer Datenbank, welche in der Literatur erwähnte funktionelle oder physikalische Interaktionen von Proteinen beinhaltet. Die Interaktionen können entweder direkt sein oder auch über weitere Gene erfolgen, die nicht in der Ergebnisliste vorhanden sind.
Die Ausgabe zeigt die Netzwerke, für die die meisten Interaktionen in der Liste gefunden wurden. Zusätzlich erhält der Anwender eine Auflistung von signifikant angereicherten funktionalen Themenkomplexen.
Hierbei werden die Kategorien

- *function and disease*
- *biofunctions*
- *canonical pathways*

unterschieden.

3.2.6 Untersuchung auf gemeinsame cis-regulatorische Motive im Promotorbereich von ko-regulierten Genen

Ein Aspekt der funktionellen Analyse einer Genliste ist die Untersuchung auf eine mögliche Regulation durch einen gemeinsamen Transkriptionsfaktor. Eine Methode zur Identifizierung von cis-regulatorischen Motiven wurde innerhalb dieser Arbeit etabliert und wird ausführlich in Kapitel 4 beschrieben.

Diese in diesem Kapitel beschriebenen Methoden werden im Folgenden auf einen Mikroarray-Datensatz angewendet, der die transkriptionelle Aktivierung und Reprimierung durch das Onkogen c-myc in einem transgenen Mausmodell untersucht. Da eine Reihe von Funktionen des Transkriptionsfaktors c-Myc bereits sehr gut charakterisiert sind, soll hiermit getestet werden, ob die Funktionen mit den Methoden gefunden werden und sie sich somit zur Hypothesengenerierung für andere, weniger gut charakterisierte Datensätze eignen.

3.3 Das Onkogen c-Myc

c-myc ist ein Proto-Onkogen, das ursprünglich als virales Onkogen (v-myc) des MC29-Stammes des *avian leukemia virus* (ALV) identifiziert wurde (Sheiness, Fanshier et al. 1978). ALV ist ein Retrovirus, das neben Karzinomen und Sarkomen die

Myelocytomatose in Vögeln induziert. Das c-myc Gen wurde erstmals 1982 im Huhn als das zelluläre Homolog von v-myc isoliert (Vennstrom, Sheiness et al. 1982). Anschließend erfolgte die Charakterisierung im Mensch, der Ratte und der Maus (Dalla-Favera, Gelmann et al. 1982). Es ist eines der am häufigsten aktivierten Onkogene. Man nimmt an, dass es in ca. 20% der humanen Tumore involviert ist. Dies ist ein Grund, warum es seit seiner Entdeckung sehr intensiv untersucht wurde.

c-Myc ist in eine Vielzahl von zellulären Prozessen involviert. Es agiert vorrangig als Transkriptionsfaktor und kann sowohl aktivierend als auch reprimierend wirken. Die transkriptionelle Aktivierung von Zielgenen durch c-Myc ist recht gut untersucht. Sie erfolgt gemeinsam mit dem bHLH-ZIP (basischer Helix-Loop-Helix-Leucin-Zipper) Protein Max als Dimerisierungspartner. Das Myc/Max-Heterodimer bindet dabei an die kanonische E-Box-Sequenz 5'-CACGTG-3' (Blackwell, Huang et al. 1993) und aktiviert die Genexpression u.a. durch die Rekrutierung von Histon-Acetyltransferasen (HAT-Komplexen) und Nukleosomen-remodullierenden -Komplexen (Cheng et al. 1999, Park et al 2001, Sommer et al 1997).

Die reprimierende Funktionsweise von c-Myc ist dagegen noch nicht genau geklärt. Sie erfolgt u.a. indirekt über die Initiator-Elemente (Inr), die zusammen mit Max und Miz1 oder Sp1 innerhalb des Promotors binden (Oster, Ho et al. 2002). Auch über die Proteine der Mad-Familie (Mad1/Mad2/Mxi1, Mad3 und Mad4) kann eine Reprimierung von Myc-Zielgenen erfolgen. Diese bilden dann einen Komplex mit Max und kompetitieren mit dem Myc/Max Komplex um die E-Box Bindungsstelle. Die Mad/Max Dimere reprimieren die Transkription, indem sie einen Komplex, der Sin3, N-CoR und HDAC1 und 2 beinhaltet, rekrutieren. Dieser veranlasst im Promotorbereich der Zielgene die Deacetylierung der Histonschwänze, was wiederum zu einer geschlossenen Chromatinkonformation führt (Ayer, Lawrence et al. 1995; Alland, Muhle et al. 1997). Ein weiterer reprimierender Mechanismus erfolgt durch die direkte Interaktion mit dem transkriptionellen Aktivator Miz-1. c-Myc bildet mit Miz-1 einen Komplex, der spezifische Miz-1 Zielgene reprimiert (Ayer, Lawrence et al. 1995; Seoane, Pouponnot et al. 2001; Wanzel, Herold et al. 2003). Auch die Interaktion von c-Myc mit den CAAT-Box bindenden Proteinen wie NF-Y wurde als möglicher Reprimierungsmechanismus beschrieben (Roy, Meisterernst et al. 1991; Seoane, Pouponnot et al. 2001; Wu, Cetinkaya et al. 2003).

Es wurden verschiedene Genexpressionsstudien durchgeführt, um die Zielgene von c-Myc zu identifizieren und damit weitere Erkenntnisse über zelluläre Prozesse, die von c-Myc reguliert werden, zu erhalten (Coller, Grandori et al. 2000; Guo, Malek et al. 2000; Schuhmacher, Kohlhuber et al. 2001; Menssen and Hermeking 2002; Watson, Oster et al. 2002; Mao, Watson et al. 2003). Diese Studien ergaben u.a. die Beteiligung der c-Myc Zielgene an verschiedenen metabolischen Prozessen, der Proteinbiosynthese, der Regulation des Zellzyklus, der Zelladhäsion und des Zytoskelettes. Weiterhin induziert c-Myc Gene, die unter Restriktion von Nährstoffen und Wachstumsfaktoren zur Apoptose führen und beeinflusst die intrazelluläre Signalweiterleitung wie z.B. bei der Angiogenese (Baudino, McKay et al. 2002).

Zellzyklus

Die ersten Erkenntnisse über die Funktionen von c-Myc waren die Fähigkeiten, die Zellteilung anzuregen und die Zelldifferenzierung zu hemmen. Zu den c-Myc Zielgenen, die hier eine Rolle spielen, gehören Cyclin D1 und D2, Cdk4 und Cyclin B1 (Hermeking, Rago et al. 2000; Bouchard, Dittrich et al. 2001; Menssen and Hermeking 2002; Fernandez, Frank et al. 2003). Cdk4 und Cdk2 werden aktiviert, indem sie mit ihren regulatorischen Untereinheiten assoziieren. Die aktivierten Komplexe CyclinD/Cdk4 und CyclinE/Cdk2 phosphorylieren dann das Retinoblastomaprotein (Rb), was wiederum zur Aktivierung des Transkriptionsfaktors E2F führt. E2F ist essentiell für die Expression von Faktoren, die für die DNA-Synthese notwendig sind (Bouchard, Thieke et al. 1999; Hermeking, Rago et al. 2000; Bouchard, Dittrich et al. 2001). Die Regulation der Cycline und CKI begünstigen, bei einer Überexpression von c-Myc, den Übergang von der G1 zur S-Phase (Adhikary and Eilers 2005).

Proteinsynthese

c-Myc spielt bei der Proteinsynthese an mehreren Stellen eine Rolle. So wird bei der Überexpression von c-Myc im Zuge der Erhöhung der Zellproliferation auch die Proteinbiosynthese verstärkt. Bei dem Vergleich von Fibroblasten, in denen c-Myc überexprimiert wurde, mit Zellen in denen c-Myc abgeschaltet war, erhöhte sich die Proteinsynthese um das 3-fache (Mateyak, Obaya et al. 1997). Die entsprechenden Gene kodieren für ribosomale RNAs und Proteine der ribosomalen Biogenese (Greasley, Bonnard et al. 2000; Kim, Li et al. 2000; Boon, Caron et al. 2001; Schlosser, Holzel et

al. 2003; Poortinga, Hannan et al. 2004). Dies erfolgt u.a. über den direkten Einfluss von c-Myc auf die RNA Polymerase III-abhängige Transkription von Genen, indem es den Pol III-spezifischen Transkriptionsfaktor TFIIIB bindet und ihn damit aktiviert (Gomez-Roman, Grandori et al. 2003). Auch für die rRNA Transkription wurde gezeigt, dass sie direkt durch c-Myc stimuliert wird (Arabi, Wu et al. 2005; Grandori, Gomez-Roman et al. 2005; Grewal, Li et al. 2005).

Metabolismen

Den Einfluss von c-Myc auf eine Reihe von Stoffwechselwegen des Zellmetabolismus konnte ebenfalls bereits gezeigt werden. Hierzu gehören der Glukosemetabolismus (Osthus, Shim et al. 2000; Menssen and Hermeking 2002; O'Connell, Cheung et al. 2003; Kim, Zeller et al. 2004), die Mitochondrienbiogenese (Wonsey, Zeller et al. 2002; Morrish, Giedt et al. 2003; O'Connell, Cheung et al. 2003; Orian, van Steensel et al. 2003), der Eisenmetabolismus (Wu, Polack et al. 1999; Bowen, Biggs et al. 2002; O'Connell, Cheung et al. 2003) und der DNA-Metabolismus (Bello-Fernandez, Packham et al. 1993; Miltenberger, Sukow et al. 1995). Für die Nukleotidsynthese relevante c-Myc Zielgene sind die Carbamoyl-Phosphate Synthase, die Aspartate Transcarbamylase und Dihydroorotase (Cad) und die Ornithine Decarboxylase (Odc).

Calcium/Calcineurin

Die Expression von c-Myc kann über einen Ca2+/Calcineurin/NF-AT abhängigen Stoffwechselweg aktiviert werden (Buchholz, Schatz et al. 2006).

Calcineurin ist eine Phosphatase, die bei der Regulation der Immunantwort eine wichtige Rolle spielt. Sie agiert u.a. über einen Calcium-Calmodulin-abhängigen Signalweg. Durch die Bindung von Calcium und Calmodulin erfolgt eine strukturelle Veränderung des Calcineurins, wodurch eine Phosphorylierungsdomäne freigelegt wird, welche dann NF-AT (nuclear factor of activated T cells) dephosphorylieren kann. Infolge der Dephosphorylierung wird NF-AT aktiviert und in den Zellkern verlagert (Crabtree and Olson 2002). Dort aktiviert es die Transkription von c-Myc durch direkte Bindung im Promotorbereich. NF-AT ist häufig in Pankreaskarzinomen überexprimiert und erhöht deutlich das maligne Potential der Tumorzellen (Buchholz, Schatz et al. 2006).

Wachstumsfaktoren

Die c-myc Expression erfolgt in Abhängigkeit von Wachstumsfaktoren. So ist sie in ruhenden, differenzierten Zellen kaum detektierbar, nach Stimulation mit Wachstumsfaktoren wird sie jedoch innerhalb von Minuten stark induziert (Bravo, Burckhardt et al. 1985). Zu den Wachstumsfaktoren, für die eine Stimulation nachgewiesen wurde, gehören die epidermalen Wachstumsfaktoren. Diese induzieren neben der c-Myc Expression auch die Stabilität bzw. Degradation von c-Myc über die Phosphorylierung der Aminosäuren Serin 62 und Threonin 58 (Alvarez, Northwood et al. 1991).

Zelladhäsion und -motilität

C-Myc Zielgene spielen ebenfalls eine Rolle bei der Zelladhäsion und –motilität. So werden eine Vielzahl von Genen, die für Proteine des Zytoskelettes und der Zelladhäsion kodieren, infolge der c-myc Überexpression herunterreguliert (Yang, Geddes et al. 1991; Yang, Gilbert et al. 1993). Diese Proteine spielen vermutlich eine Rolle bei der c-Myc abhängigen neoplastischen Transformation von Zelllinien, die von einem nicht adhärenten Wachstum begleitet wird. Nach Überexpression von c-myc in den Keratinozyten transgener Mäuse wurden 218 differentiell exprimierte Gene identifiziert (Frye, Gardner et al. 2003). Ein Drittel dieser Gene sind an der Zelladhäsion beteiligt und 11% kodieren für Zytoskelettproteine. Die veränderte Expression der Zytoskelettgene beeinflusst die Zellmorphologie. So haben c-myc Null-Fibroblasten deutlich mehr Actin-Stress-Fasern und fokale Adhäsionsproteine im Vergleich zu den Fibroblasten, die ein rekonstituiertes c-myc Gen haben (Shiio, Donohoe et al. 2002). Aktin-Stress-Fasern sind Aktinbündel, die für die Adhäsion und den Zusammenhalt der Fibroblasten wichtig sind. Zu den Zytosklettproteinen, die in Fibroblasten mit einem rekonstituierten c-Myc herunterreguliert werden, gehören Aktin und Cdc42. Cdc42, das zur Familie der Rho-GTPasen gehört, reguliert die Ausbildung von Filopodien. Filopodien sind Plasmaausstülpungen, die mittels Bündeln von Aktinfilamenten gebildet werden und insbesondere zur Zellmigration dienen. Weiterhin wurde die Beteiligung von c-myc bei der Bewegung von hämapoetischen Stammzellen aus der Stammzellnische gezeigt (Wilson, Murphy et al. 2004).

3.4 Untersuchung eines cDNA Mikroarray Datensatzes auf funktionelle Gemeinsamkeiten

3.4.1 Datensatz

Als Datensatz wurden die Ergebnisse eines Mikroarrayversuches gewählt, bei dem die Genexpression der T-Lymphozyten transgener Mäuse miteinander verglichen wurde. Diese Mäuse exprimieren entweder nur einen tetracyclin-abhängigen transkriptionellen Aktivator eines Eμ-SRalpha Promotor, oder auch als zweites Transgen ein humanes c-myc Wildtyp-Allel, welches von einem Minimalpromotor mit Tet-Operator initiiert wird. Der Tet-Operator wird durch die gleichzeitige Expression des tetracyclin-abhängigen transkriptionellen Aktivators in T-Zellen aktiviert und bewirkt somit eine T-Lymphozyten spezifische Expression von c-myc. Die Mäuse beider Gruppen wurden im Alter von 5 Wochen getötet, und die RNA aus den T-Lymphozyten präpariert (s. Abbildung 13). Der Versuch wurde von Tobias Otto am Institut für Molekularbiologie und Tumorforschung, (IMT) in Marburg durchgeführt.

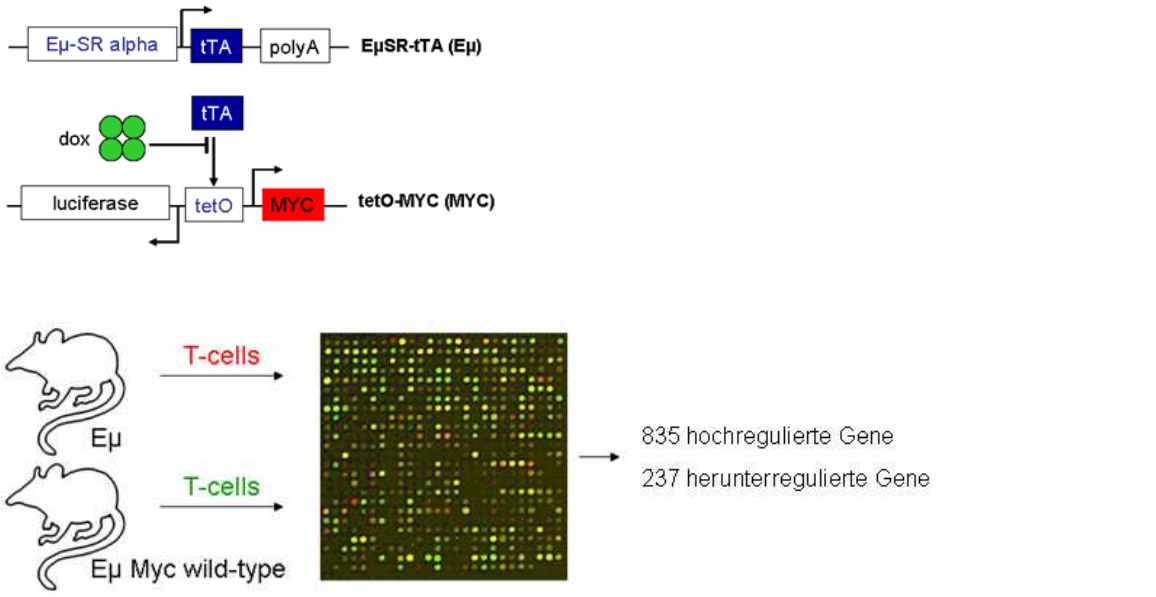


Abb. 13: Genexpressionsversuch, zum Vergleich der c-Myc abhängigen Genexpression in T-Lymphozyten von transgenen Mäusen. (a) Für diesen Versuch wurden Lymphozyten von 5 Wochen alten Mäusen verwendet, die entweder nur einen tetracyclin-abhängigen transkriptionellen Aktivator eines Eμ-SRalpha Promotors, oder auch als zweites Transgen ein humanes c-myc Wildtyp-Allel, welches von einem Minimalpromotor mit Tet-Operator initiiert wird, transkribieren. Der Tet-Operator wird durch die gleichzeitige Expression des tetracyclin-abhängigen transkriptionellen Aktivators in T-Zellen aktiviert und bewirkt somit eine T-Lymphozyten spezifische Expression von c-myc. (b) Die T-Zellen der beiden transgenen Maustypen wurden in einem Mikroarray-Versuch verwendet. Das Experiment wurde als Referenzdesign durchgeführt. Insgesamt sind 1072 Gene differentiell exprimiert davon 835 Gene hoch- und 237 Gene herunterreguliert.

Die in diesem Versuch verwendeten Mikroarrays wurden am IMT hergestellt. Es handelt sich um 21k cDNA-Mikroarrays, für dessen Herstellung zwei vom *National Institute on Aging* (NIA, USA) stammende Klon-Bibliotheken (15k und 7,4k cDNA Klonset) verwendet wurden.

3.4.2 Statistische Auswertung

Als Versuchsdesign wird das Referenzdesign gewählt, da zwei Gruppen mit jeweils drei biologischen Proben miteinander verglichen werden. Als Referenzprobe wird eine Mischung aller Proben verwendet. Zur Normalisierung wird eine *Printtip-Lowess* Normalisierung jedes Mikroarrays durchgeführt. Die Selektion differentiell exprimierter Gene erfolgt mittels der SAM-Methode (Tusher, Tibshirani et al. 2001). Bei einer *false discovery rate* <5% sind insgesamt 1072 Gene signifikant unterschiedlich zwischen beiden Phänotypen reguliert. Davon sind 835 Gene hoch- und 237 Gene herunter-reguliert.

3.4.3 Funktionelle Interpretation des Datensatzes

Für das Gen c-myc wurden bereits zahlreiche Zielgene beschrieben, die bestimmten funktionellen Bereichen zugeordnet werden konnten. Im Folgenden wird getestet, ob diese Themenkomplexe mit den dargestellten Methoden gefunden werden und sie sich somit zur Charakterisierung bzw. Hypothesengenerierung von Datensätze eignen.

Expression Analysis Systematic Explorer (EASE)

Die Ergebnisse der Auswertung sind in 10.1 im Anhang aufgelistet. Es sind 101 Kategorien signifikant angereichert, die jedoch, entsprechend der Strukturierung der *gene ontology terms*, eine sehr hohe Redundanz aufweisen. Es finden sich innerhalb der Liste signifikant angereicherter funktioneller Themenbereiche eine Reihe der bereits bekannten mit c-Myc assoziierten Funktionen wieder. So beinhaltet die am stärksten angereicherte Kategorie Gene, die an der Zusammensetzung der Ribosomen, der Ribosomenbiogenese, der rRNA Prozessierung und dem rRNA Metabolismus beteiligt sind. Eine weitere bereits beschriebene biologische Funktion, für die Gene in der Ergebnisliste angereichert sind, ist der Nukleinsäure-Metabolismus. Dagegen treten einige Themen, die man im Zusammenhang mit dem experimentellen Phänotyp erwarten würde, wie z.B. die Zellteilung oder die Apoptose, nicht signifikant angereichert auf.

Gene Set Enrichment Analysis (GSEA)

Die *Gene Set Enrichment Analysis* (GSEA) ist eine Methode, die untersucht, ob für bestimmte im Vorfeld festgelegte Gensets statistisch signifikante Unterschiede zwischen verschiedenen Phänotypen bestehen. Es werden zwei Listen generiert. Eine Liste, die Gensets beinhaltet, in denen Gene angereichert sind, die durch c-Myc hochreguliert werden. Die andere Liste beinhaltet Gensets, in denen Gene angereichert sind, die durch c-Myc herunterreguliert werden.

Gensets, in denen Gene angereichert sind, die durch c-Myc hochreguliert werden

In Tabelle 2 sind die 30 Gensets, die die größte Anreicherung von Genen aufweisen, die durch c-Myc hochreguliert werden. Die vollständige Liste der Gensets, die eine signifikante Anreicherung zeigen, ist in einer Tabelle im Anhang (s.10.2) aufgelistet.
In der Liste der Gensets, in denen Gene angereichert sind, die durch c-Myc hochreguliert werden, finden sich hauptsächlich funktionelle Bereiche, die bereits im Zusammenhang mit c-Myc beschrieben wurden und somit erwartet werden. So ist auf Rang 13 ein Genset, das ribosomale Proteine beinhaltet, die an der Proteinsynthese beteiligt sind. Auf Rang 15, 17 und 20 der Liste treten Gensets auf, die Gene beinhalten, die im Promotorbereich ein E-Box Bindungsmotiv (V$MYCMAX_01, V$USF_C und V$NMYC_01) haben, somit auf die Funktion von c-myc als transkriptioneller Aktivator deuten. Verschiedene Gensets der Liste beinhalten Gene, die c-Myc abhängig reguliert wurden. So ist auf Rang 18 ein Genset, das Gene beinhaltet, von denen in mindestens 3 Publikationen beschrieben worden ist, dass sie durch c-Myc hochreguliert werden (Zeller, Jegga et al. 2003), auf Rang 29 ein Genset mit Genen, die in Hepatomgewebe von transgenen c-myc-Mäusen herauf reguliert sind. Auf Position 23 befindet sich ein Genset, das c-Myc-abhängige Gene, die in verschiedenen Systemen beschrieben wurden, beinhaltet (Zeller, Jegga et al. 2003).
Die meisten der weiteren Gensets beinhalten Gene, die in verschiedenen Tumoren mit bekannten Onkogenen ko-reguliert auftreten. Hierzu gehören Fbl, Tpt1, Npm1, Eif3s6, St13, Actg1, Casp8ap2, JunD, Dap3 und Nme2. Einige dieser Gene wurden bereits im Kontext mit c-myc beschrieben, ihr Auftreten bestätigt somit auch bereits bekannte Ergebnisse. So ist Fbl ein bekanntes c-myc Zielgen, welches bei der Überexpression von c-myc hochreguliert wird (Robanus-Maandag, Bosch et al. 2003). Fbl ist eine Komponente des snRNP (*small nuclear ribonucleoproteins*) von dem angenommen

wird, dass es in den ersten Schritten der Prozessierung der pre-ribosomalen RNA beteiligt ist, Die Transkription von Nucleophosmin/B23 (Npm1) wird durch c-Myc während des Zellzyklus aktiviert (Yeh, Huang et al. 2006). ST13 kodiert für das Hsp70 interagierende Protein (HIP), ein Ko-faktor des 70-kDa Hitzeschockproteins (Hsc/Hsp70). Hsp70 ist mit c-Myc im Zellkern ko-lokalisiert, es hat die Fähigkeit, freie Level von c-Myc in der Zelle zu reduzieren (Jaattela, Wissing et al. 1992). Beim Vergleich dieser Gensets fällt auf, dass alle die gleiche Gruppe ribosomaler Proteine enthalten. Sie sind maßgeblich daran beteiligt, dass diese Gensets innerhalb der Ergebnisliste auftreten. Dies erklärt das Auftreten dieser Gensets ohne das bisher ein direkter Zusammenhang mit c-myc beschreiben wurde.

Tab. 2: Liste der 30 Gensets, die die größte Anreicherung von Genen aufweist, die durch c-Myc hochreguliert werden. In den Spalten sind von links nach rechts der Name des Gensets, die Größe des Gensets, der *enrichment score* (ES), der auf die Anzahl der Gene des Gensets normalisierte *enrichment score* (NES), der nominale p-Wert und die *false discovery rate* (FDR q-value).

Rang	NAME	SIZE	ES	NES	NOM p-val	FDR q-val
1	GNF2_FBL	102	-0.7001	-2.2207	0	0
2	MORF_TPT1	69	-0.7289	-2.1744	0	0
3	MORF_ACTG1	96	-0.6402	-2.0798	0	0.00046
4	GNF2_ST13	48	-0.796	-2.0884	0	0.00053
5	MORF_NPM1	113	-0.6553	-2.1208	0	0.00061
6	GNF2_EIF3S6	84	-0.77	-2.0999	0	0.00064
7	GCM_NPM1	83	-0.6957	-2.1096	0	0.0008
8	FLOTHO_CASP8AP2_MRD_DIFF	40	-0.5554	-2.0488	0	0.00112
9	MORF_JUND	47	-0.6059	-2.0437	0	0.00172
10	GNF2_NPM1	44	-0.6949	-2.0343	0	0.00188
11	GNF2_DAP3	88	-0.6059	-2.0262	0	0.00189
12	MORF_NME2	108	-0.6034	-2.0138	0	0.00199
13	RIBOSOMAL_PROTEINS	60	-0.7807	-2.0095	0	0.00207
14	GCM_TPT1	45	-0.8338	-1.9944	0	0.00257
15	V$MYCMAX_01	133	-0.5495	-1.9995	0	0.00275
16	GNF2_RBBP6	47	-0.5614	-1.9744	0	0.00326
17	V$USF_C	140	-0.4808	-1.9752	0	0.00345
18	ZELLER_MYC_UP	18	-0.7532	-1.982	0	0.00348
19	MORF_DEAF1	41	-0.6887	-1.9623	0	0.00433
20	V$NMYC_01	133	-0.4631	-1.9493	0	0.00494
21	MORF_FBL	103	-0.601	-1.9402	0	0.00586
22	GNF2_GLTSCR2	25	-0.8555	-1.9311	0	0.00655
23	MYC_TARGETS	32	-0.7545	-1.9271	0	0.00691
24	AGUIRRE_PANCREAS_CHR17	30	-0.5466	-1.916	0	0.00809
25	SHEPARD_CRASH_AND_BURN_MUT_VS_WT_UP	90	-0.4857	-1.8976	0	0.00948
26	HESS_HOXAANMEIS1_DN	49	-0.5975	-1.8923	0.0039	0.00976
27	HESS_HOXAANMEIS1_UP	49	-0.5975	-1.8923	0.0039	0.01012
28	GCM_PSME1	68	-0.5532	-1.8935	0	0.01014
29	SCHUMACHER_MYC_UP	34	-0.6739	-1.8776	0	0.01137

Gensets, in denen Gene angereichert sind, die durch c-Myc herunterreguliert werden

In Tabelle 3 sind die 30 Gensets, die die größte Anreicherung von Genen aufweist, die durch c-Myc herunterreguliert werden. Die vollständige Liste ist in einer Tabelle im Anhang (s.10.3.) aufgelistet. Auch diese Liste beinhaltet Gensets, die anhand der bereits bekannten Funktionen von c-myc erwartet werden können, es finden sich hier aber auch Gensets, die die Erstellung neuer Hypothesen ermöglichen.

Die größte Anreicherung zeigt ein Genset, mit Genen, die mit CD97 korrelieren. Da CD97 bereits als c-Myc Zielgen beschrieben wurde (Li, Van Calcar et al. 2003), ist das Auftreten im Rahmen dieser Auswertung zu erwarten. CD97 ist Mitglied der EGF-TM7 Familie, tritt an der Oberfläche von aktivierten Leukozyten auf und enthält zumeist Zelladhäsionsmotive. Gene, die in der Zelladhäsion und –motilität eine Rolle spielen, wurden bereits als c-Myc Zielgene beschrieben. Auf Rang 4 ist das Genset GNF2_RAP1B, das eine große Übereinstimmung mit der CD97 Genliste aufweist.

Auf Rang 3 ist ebenfalls ein Genset, welches im Rahmen der Auswertung zu erwarten wäre. Es handelt sich um Gene, die Aufgaben bei der Zellbewegung übernehmen. Die Überexpression von c-myc führt zu einer Herunterregulation von Genen, die für Proteine des Zytosklettes und der Zelladhäsion kodieren. Somit ist das Auftreten dieses Gensets in Korrelation mit dem Phänotyp der Eμ-Maus zu erwarten. Neben CD97 finden sich z.B. Moesin und WASL (*Wiskott-Aldrich syndrome-like*) in diesem Genset, beides Proteine, welche als Verbindungselement zwischen integralen Membranproteinen und dem Aktinzytoskelett fungieren.

Auf Rang 9 ist ein Genset, bestehend aus Genen des Insulin-Signalweges. Kaneto, Sharma et al. (2002) zeigten, dass c-Myc in HeLa und HepG2- Zellen die Transkription von Genen des Insulin-Signalweges unterdrückt, indem es die NeuroD-vermittelte transkriptionelle Aktivierung hemmt. NeuroD ist ein Transkriptionsfaktor, der die Expression der Insulingene reguliert. Es gehört wie c-Myc auch zur Familie der Helix-loop-Helix (bHLH) Transkriptionsfaktoren und bindet als Heterodimer mit anderen bHLH Proteinen an die E-Box. Mutationen dieses Genes führen zu Diabetes mellitus Typ II. Das Auftreten des Gensets bei denen Gene, die durch die Aktivierung von c-Myc herunterreguliert werden, bestätigen dieses Ergebnis.

Andere Gensets ermöglichen die Generierung neuer Hypothesen. So ist auf Rang 2 der Liste ein Genset, das Mausgene beinhaltet, die im Zusammenhang mit Calcium-, Calcineurin- bzw. NF-AT-abhängiger Signaltransduktion stehen. c-myc wurde bereits

als Zielgen von NF-AT beschrieben. Die Aktivierung der Transkription von c-myc durch NF-AT erfolgt durch eine Ca^{2+}/Calcineurin abhängige Signaltransduktion (Buchholz et al., 2006).

Tab. 3: Die Liste der 30 Gensets, die die größte Anreicherung von Genen aufweist, die durch c-Myc herunterreguliert werden. In den Spalten ist von links nach rechts der Name des Gensets, die Größe des Gensets, der *enrichment score* (ES), der auf die Anzahl der Gene des Gensets normalisierte *enrichment score* (NES), der nominale p-Wert und die *false discovery rate* (FDR q-value).

NAME	SIZE	ES	NES	NOM p-val	FDR q-val
GNF2_CD97	22	0.7801	2.1302	0	0.0083
CALCINEURIN_NF_AT_SIGNALING	47	0.5810	2.0952	0	0.0063
CELL_MOTILITY	64	0.5667	2.0639	0	0.0064
GNF2_RAP1B	25	0.7377	2.0202	0.003571429	0.0141
EGFPATHWAY	22	0.6861	2.0200	0	0.0113
BASSO_GERMINAL_CENTER_CD40_UP	50	0.6663	2.0079	0	0.0118
GLEEVECPATHWAY	17	0.6647	1.9974	0	0.0142
GNF2_ITGB2	31	0.6402	1.9973	0	0.0124
INSULINPATHWAY	17	0.6506	1.9629	0	0.0181
CARIES_PULP_UP	94	0.5631	1.9598	0.003610108	0.0164
TCRPATHWAY	34	0.6633	1.9566	0	0.0153
GNF2_ITGAL	27	0.7408	1.9564	0	0.0140
PASSERINI_ADHESION	22	0.6549	1.9515	0	0.0140
YU_CMYC_DN	32	0.6308	1.9433	0.003717472	0.0146
GNF2_PTPRC	35	0.7020	1.9407	0	0.0149
METPATHWAY	24	0.5973	1.9340	0	0.0149
BCRPATHWAY	27	0.6167	1.9267	0.003623189	0.0175
HALMOS_CEBP_UP	22	0.5661	1.9045	0	0.0213
GNF2_PECAM1	21	0.7129	1.9011	0	0.0213
GNF2_CASP1	45	0.5983	1.9011	0	0.0202
ASTON_DEPRESSION_UP	22	0.6419	1.9006	0	0.0197
PYK2PATHWAY	23	0.5916	1.8980	0.003649635	0.0198
GNF2_SELL	26	0.7242	1.8855	0	0.0222
PDGFPATHWAY	23	0.6473	1.8787	0	0.0235
FCER1PATHWAY	31	0.5723	1.8720	0	0.0243
CHIARETTI_T_ALL	121	0.5034	1.8633	0	0.0264
AT1RPATHWAY	28	0.5439	1.8593	0.003333333	0.0267
GNF2_CASP8	16	0.7627	1.8519	0.003597122	0.0288
BIOPEPTIDESPATHWAY	29	0.5785	1.8491	0	0.0301

Zu den Genen dieses Gensets, die innerhalb des Datensatzes durch c-Myc herunterreguliert werden, gehören ITK (IL-2 induzierbare T-Zell spezifische Tyrosinkinase) und MAPK8 (JNK). Beide Gene spielen während der Lymphozyten-Aktivierung eine Rolle. ITK gehört zu der Tec Familie, einer Gruppe von Tyrosin-Kinasen (Schwartzberg, Finkelstein et al. 2005). Die Tec Kinasen Rlk and Itk aktivieren innerhalb der Signalkaskade unter anderem die Transkription von NF-AT. Auch MAPK8 ist in diese

Signalkaskade involviert und führt ebenfalls zur Aktivierung von NF-AT. Da die NF-AT-regulierenden Gene durch c-Myc herunterreguliert werden, lässt sich hieraus eine mögliche Regulation dieser Gene über eine Rückkopplungshemmung durch c-Myc schließen.

Auf Rang 5 der Liste ist ein Genset, das 22 Gene des EGF-Signalweges beinhaltet. Die Gene STAT1, MAPK8, PIK3R1 JUN, EGFR, MAP2K1, RASA1, GRB2 und PRKCB1 dieses Pathways sind infolge der c-Myc Induktion herunterreguliert. Auch hier liegt möglicherweise, wie bei der Calcium/Calcineurin/NF-AT-abhängigen Signaltransduktion, eine Rückkopplungshemmung vor.

Connectivity Map

Mit der *Connectivity Map* Software können molekulare Zusammenhänge zwischen den Ergebnissen von Genexpressionsversuchen und der Wirkung kleiner bioaktiver Moleküle hergestellt werden.

Die folgende Tabelle zeigt die bioaktiven Moleküle, die eine signifikante Korrelation mit den differentiell exprimierten Genen der Ergebnisliste haben. Dies kann eine positive oder eine negative Korrelation sein.

Tab. 4: Bioaktive Moleküle, die eine signifikante Korrelation mit den differentiell regulierten Genen infolge c-Myc Induktion aufzeigen. Ein negativer Wert in der Spalte „Anreicherung“ zeigt eine Korrelation mit den Genen auf, die durch c-Myc herunterreguliert wurden, ein positiver Wert eine Korrelation mit Genen, die durch c-Myc hochreguliert werden.

Rang	Name des bioaktiven Moleküls	Anreicherung	p-Wert
1	trichostatin A	-0.613	0
2	5182598	0.945	0.0064
3	pyrvinium	-0.914	0.0143
4	sirolimus	-0.456	0.0217
5	fluphenazine	-0.678	0.0227
6	imatinib	0.874	0.0302
7	vorinostat	-0.865	0.0349
8	genistein	0.5	0.0357
9	calmidazolium	-0.852	0.0419
10	copper sulfate	0.635	0.0426
11	NU-1025	0.845	0.0459

Trichostatin A/Vorinostat

Trichostatin A und Vorinostat sind beides Histon-Deacetylase-Hemmer. Gene, die in Zelllinien, die mit den beiden Substanzen behandelt wurden, hochreguliert wurden,

korrelieren mit den Genen, die infolge c-myc Überexpression herunterreguliert werden. Sowohl für Trichostatin A als auch für Vorinostat wurde gezeigt, dass sie c-myc herunterregulieren (Li and Wu 2004; Kumagai, Wakimoto et al. 2007), somit ist eine solche Korrelation zu erwarten.

Sirolimus/Rapamycin

Eine negative Korrelation findet sich auch mit dem makrozyklischem Immunsuppressivum Sirolimus, welches auch unter dem Namen Rapamycin bekannt ist. Sirolimus ist ein Makrolidantibiotikum aus Streptomyceten (*Streptomyces hygroscopicus*), das einen antiproliferativen Effekt auf Zellen hat. Über die Komplexbildung mit mTOR (*mammalian Target of Rapamycin*) inhibiert es die cap-abhängige Proteinsynthese und blockiert die Biosynthese von Komponenten der Ribosomenbiosynthese, was zu einem G_1-Zellzyklusarrest führt. Die Blockade der cap-abhängigen Proteinsynthese führt ebenfalls dazu, dass die translationale Komponente der durch c-Myc bedingten Aktivierung nicht erfolgt (West, Stoneley et al. 1998). Dementsprechend findet sich eine Korrelation mit Genen, die in den c-Myc-Null-Mäusen stärker exprimiert sind.

Genistein

Genistein ist ein Isoflavon der Sojabohne, welches einen antiproliferativen Effekt auf eine Reihe von Tumorzellen hat. Es wurde gezeigt, dass im Prostatatumor und in den Kolontumorzelllinien HCT8 and SW837 die Behandlung von Zellen mit Genistein zu einer Reduzierung der c-Myc Expression und Translation führt (Heruth, Wetmore et al. 1995; Jagadeesh, Kyo et al. 2006). Somit würde man eine Korrelation mit den Genen, die in den Myc-Null-Mäusen stärker exprimiert sind, erwarten. Es tritt jedoch eine Korrelation mit den Genen auf, die in den Eμ-MycWT-Mäusen hochreguliert sind und somit der umgekehrte Effekt.

Fluphenazin/Calmidazolium

Fluphenazin und Calmidazolium gehören zur Gruppe der Calmodulin-Inhibitoren. Die Expressionsmuster der mit diesen Substanzen behandelten Zellen korrelieren mit Genen, die in den Myc-Null-Mäusen höher exprimiert sind als in den Eμ-MycWT-Mäusen. Dies ist ebenfalls zu erwarten, da die Calmodulin-Inhibitoren Calcium aus der

Zelle transportieren und so den Ca^{2+}/Calcineurin NF-AT-abhängigen Stoffwechselweg hemmen, der bei Anwesenheit von Calcium aktivierend auf die Transkription von c-myc wirkt (Buchholz, Schatz et al. 2006).

Ingenuity Pathway Analysis

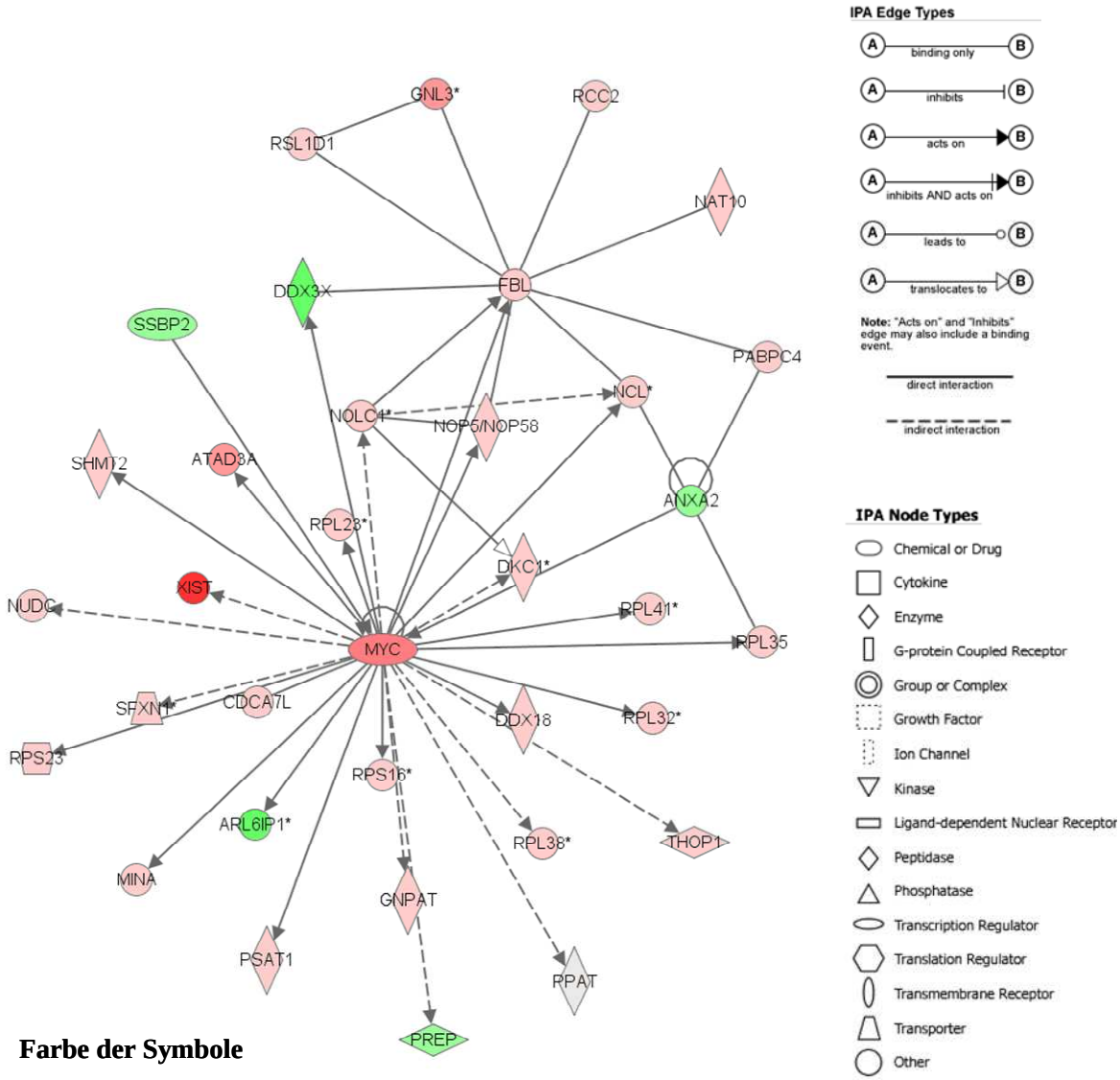
Ingenuity Pathway Analysis berechnet aus den Genen der Ergebnisliste Netzwerke anhand einer Datenbank, die bereits publizierte Interaktionen enthält. Die Reihenfolge der generierten Netzwerke entspricht der Anzahl der Interaktionen, d.h. je mehr Interaktionspartner innerhalb der Ergebnislisten gefunden wurden, desto weiter oben tritt das Netzwerk in der Liste auf.

Tab. 5: Die 10 Netzwerke mit der höchsten Zahl von Interaktionen. Der Wert „Score" ins Spalte 2 gibt einen Wert für die Robustheit des Netzwerkes, „*Focus Molecules*" für die Anzahl der berücksichtigten Interaktionen an. Im der Spalte „*Top functions*" werden die funktionelle Themenbereiche angegeben, die jedem Netzwerk anhand der Gene zugeordnet werden.

ID	Score	Focus Molecules	Top Functions
1	57	35	RNA Post-Transcriptional Modification, Protein Synthesis, Cancer
2	41	29	Cancer, Cell Death, Cardiac Dysfunction
3	39	28	Cellular Growth and Proliferation, Hematological System Development and Function, Cancer
4	39	28	Cell Cycle, Gene Expression, Cancer
5	37	27	Protein Synthesis, Cellular Growth and Proliferation, DNA Replication, Recombination, and Repair
6	35	26	Cancer, Cellular Growth and Proliferation, Connective Tissue Development and Function
7	35	26	Cellular Assembly and Organization, Cell Morphology, Cellular Development
8	21	19	Protein Synthesis, Cancer, Cell Death
9	21	19	Cell Cycle, Skeletal and Muscular System Development and Function, Cellular Development
10	20	18	Gastrointestinal Disease, Hepatic System Disease, Cell Signaling

Tabelle 5 zeigt die ersten 10 Netzwerke der Ergebnisliste, d.h. die Netzwerke, die die meisten Interaktionen aufweisen. Jedem Netzwerk werden anhand der Gene funktionelle Themenbereiche zugeordnet, welche in der Spalte „*Top functions*" aufgelistet sind. Innerhalb dieser Liste finden sich einige c-myc relevante Themenbereiche wie z.B. *protein synthesis*, *cancer*, *cell death*, *cell cycle* und *cell morphology*.

Im Folgenden werden zwei Netzwerke näher beschrieben. Das Netzwerk 1 ist in Abbildung 14 dargestellt. In diesem Netzwerk, mit insgesamt 35 Genen, sind eine Reihe von Genen dargestellt, die als direkte Interaktionpartner von c-Myc bereits beschrieben worden sind. Die Mehrheit der Gene des Netzwerkes werden durch c-Myc aktiviert.



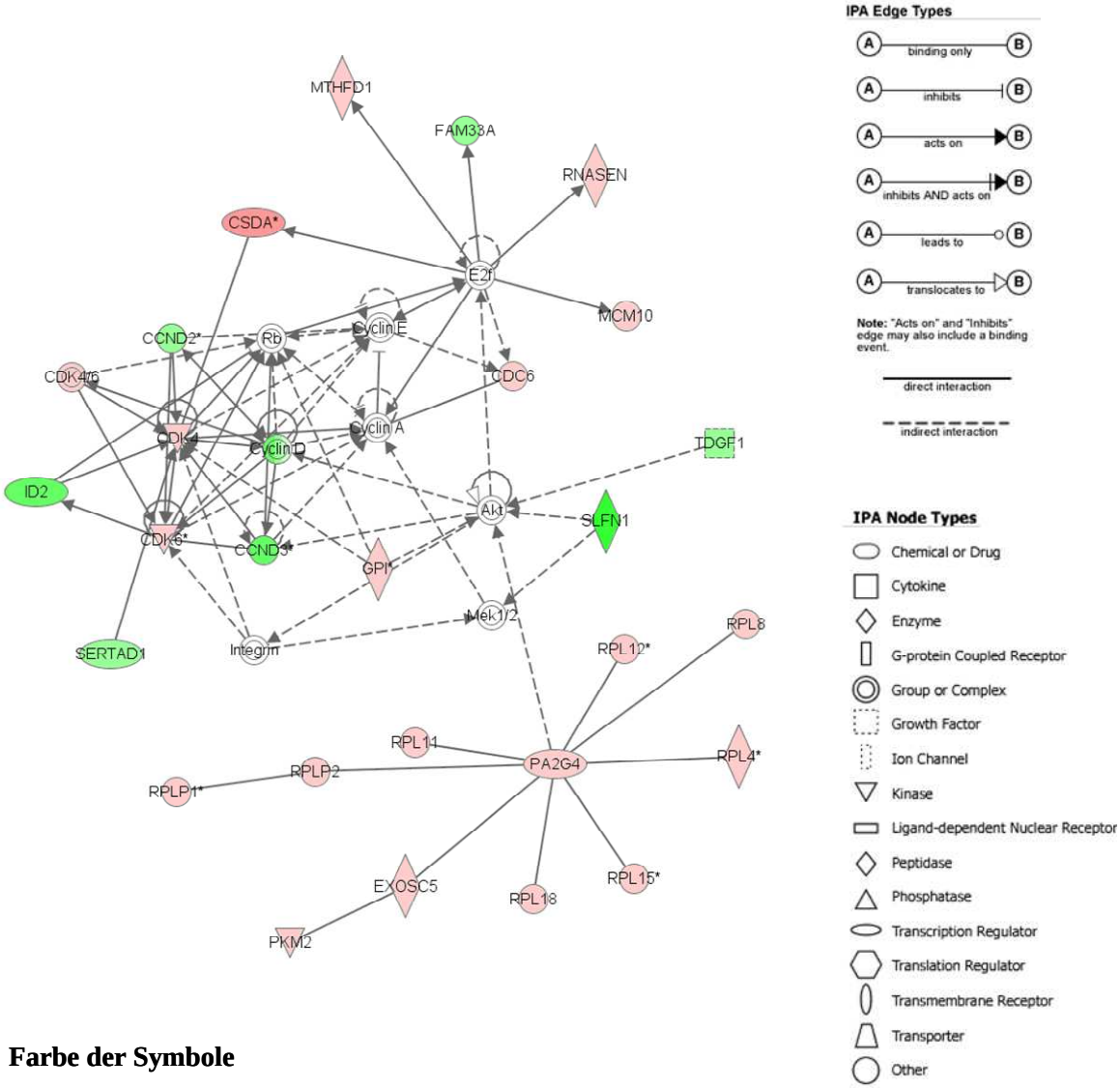
Farbe der Symbole

Rot	Gene der Ergebnisliste, welche hochreguliert sind. Die Intensität gibt die Stärke der Regulation an
Grün	Gene der Ergebnisliste, welche heruntereguliert sind. Die Intensität gibt die Stärke der Regulation an
Weiß	Gene, die nicht in der Ergebnisliste, aber im Netzwerk enthalten sind aufgrund von Interaktion mit anderen Genen.

Abb. 14: Abbildung des Netzwerkes 1, welches die meisten Interaktionen aufweist

Diese Gene werden den Themen *RNA post-transcriptional modification*, *protein synthesis* und *cancer* zugeordnet. Zu dem funktionellen Bereich der *RNA post-transcriptional modification* und *protein synthesis* gehören z.B. die ribosomalen Proteine RPL23, RPL32, RPL35, RPL38, RPL41, RPS16, RPS23 und RSL1D1, deren Transkription von c-Myc aktiviert wird. Die Gene MINA, PSAT1, PPAT und

CDCA7L, welche ebenfalls durch c-myc aktiviert werden, werden dem Bereich „*cancer*“ zugeordnet.



Farbe der Symbole

Rot	Gene der Ergebnisliste, welche hochreguliert sind. Die Intensität gibt die Stärke der Regulation an
Grün	Gene der Ergebnisliste, welche heruntereguliert sind. Die Intensität gibt die Stärke der Regulation an
Weiß	Gene, die nicht in der Ergebnisliste, aber im Netzwerk enthalten sind aufgrund von Interaktion mit anderen Genen.

Abb. 15: Darstellung des Netzwerkes 6

Dem Netzwerk 6 werden die funktionellen Bereiche *cancer*, *cellular growth and proliferation*, und *connective tissue development and function* zugeordnet.

Im unteren Bereich des Netzwerkes sich verschiedene strukturelle ribosomale Gene z.B. RPL18, RPL15, RPLP1, RPLP2, RPL4, RPL12, RPL8 und RPL11. Als zentrales Gen, welches mit verschiedenen ribosomalen Proteinen interagiert ist PA2G4, ein Gen welches Teil des prä-ribosomalen Ribonukleoproteinkomplex ist und an der Erstellung und Regulation der intermediären und späteren Schritte der rRNA Prozessierung beteiligt ist. Die ribosomalen Gene werden dem funktionellen Bereich der Proteinsynthese zugeordnet, dementsprechend wäre zu erwarten gewesen, dass diese Kategorie auch diesem Netzwerk zugeordnet würde.

Im oberen Bereich des Netzwerkes sind verschiedene Gene der Zellzyklusregulation, die eine Vielzahl von wechselseitigen Interaktionen aufweisen. Ein Teil dieser Gene sind spezifisch für das Genexpressionsprofil, welches in Säugerzellen bei der Progression aus der G1 in die S-Phase auftritt. Hierzu gehören CSDA (*cold shock domain protein A*), CCND2, welches einen Komplex mit CDK4 und CDK6 bildet, die ebenfalls in dem Netzwerk auftreten, und Cyclin D (Sherr and Roberts 1999; Donjerkovic and Scott 2000).

Neben der Netzwerkgenerierung ermöglicht *Ingenuity Pathway Analysis* auch funktionelle Klassifizierungen der Gene. Die Analyse dieses Datensatzes hinsichtlich *function and disease* ergibt eine Liste von 776 signifikant auftretenden funktionellen Bereichen. Diese sehr hohe Zahl kommt dadurch zustande, dass viele Themenbereiche redundant auftreten. So wird z.B. zwischen *cell death of lymphocytes, cell death of mononuclear leukocytes, cell death of leukocytes, cell death of blood cells* und *cell death of lymphatic system cells* unterschieden. In der Ergebnisliste treten typische funktionelle Bereiche auf, für die die c-Myc abhängige Regulation gezeigt wurde wie z.B. Apoptose, Zellteilung, Zellwachstum, EGF-Signaltransduktion, Tumorentstehung, G1/S Phasen Transition und Zellbewegung. Eine allgemeinere Klassifizierung erfolgt in der Untersuchung der Liste auf angereicherte *biofunctions*. Auch diese Liste beinhaltet 60 verschiedene signifikant angereicherte Funktionen. Die am stärksten angereicherten Themen sind *cell death, cellular growth and proliferation, cell signaling, protein synthesis, cancer und cell cycle*, somit die für c-Myc typischen funktionellen Bereiche.

Identifizierung von signifikant angereicherten cis-regulatorischen Motiven in den Promotorsequenzen der ko-regulierten Gene

Die Promotorsequenzen der differentiell regulierten Gene wurden auf gemeinsame cis-regulatorische Elemente untersucht. Eine genaue Beschreibung der Methode und die Erläuterung der Ergebnisse ist in Kapitel 4 gegeben. Zusammenfassend kann man sagen, dass cis-regulatorische Motive gefunden wurden, die auf eine c-Myc abhängige Regulation hinweisen, bzw. Bindungsstellen für andere Transkriptionsfaktoren, z.B. YY1 für die bereits eine Ko-regulation mit c-Myc beschrieben wurde.

3.4.4 Diskussion

Anhand der Ergebnisse dieses Datensatzes konnte gezeigt werden, dass die hier verwendeten bioinformatischen Methoden für die funktionelle Interpretation von Genexpressionsprofilen geeignet sind. So zeigen alle Methoden Funktionen, Interaktionen bzw. Motive auf, die bereits im Zusammenhang mit c-Myc beschrieben wurden. Neben bereits bekannten Funktionen, wie z.B. der Ribosomenbiogenese oder der Regulation der DNA-Synthese, treten auch mögliche neue Funktionen von c-Myc auf, wozu die Rückkopplungshemmung der NF-AT- und der EGF-abhängigen Genregulation gehören. Nicht alle Funktionen, die im Zusammenhang mit c-Myc zu erwarten sind, werden mit allen hier beschriebenen Methoden gefunden. So fehlen sowohl bei der EASE- als auch der GSEA-Methode die Kategorien Apoptose und Zellteilung, treten aber in der Ergebnisliste von *Ingenuity Pathway Analysis* auf. Einige Ergebnisse erscheinen im Vergleich mit der Literatur als falsch positiv, so z.B. die Korrelation der Genistein regulierten Gene mit den in den Eμ-Myc Wildtyp-Mäusen hochregulierten Genen.

Aufgrund der unterschiedlichen methodischen Ansätze ergänzen sich die verschiedenen Methoden hinsichtlich ihrer Aussage. So entwickelt *Ingenuity Pathway Analysis*, neben der funktionellen Kategorisierung, *de novo* Netzwerke anhand in der Literatur beschriebener Interaktionen. Dies erleichtert die Auffindung möglicher direkter funktioneller Zusammenhänge einzelner bzw. mehrerer Gene des Datensatzes und die Hypothesengenerierung hinsichtlich neuer Stoffwechselwege. GSEA verwendet bei der Auswertung, im Vergleich zu den anderen Methoden, die „nur“ die Listen differentiell exprimierter Gene analysieren, das gesamte Genexpressionsprofil und kann so einen

höheren Informationsgehalt ausnutzen. Diese Methode kategorisiert, wie die anderen Methoden auch, nach funktionellen Bereichen, ermöglicht aber zusätzlich den Vergleich des zu untersuchenden Genexpressionsprofils mit den Ergebnissen anderer Genexpressionsstudien. Diese Listen werden von der Software bereitgestellt, können aber auch selber erstellt werden. Auch die EASE-Methode ermöglicht die Erstellung von eigenen funktionellen Gruppen, jedoch mit der Einschränkung, dass nur auf die jeweiligen Ergebnislisten fokussiert wird.
Connectivity Map unterscheidet sich inhaltlich von den anderen Methoden, da sie die Korrelation mit der Wirkung von bioaktiven Molekülen untersucht. Das Auffinden der Korrelation der Wirkung der Histon-Desacetylase-Hemmer Trichostatin A und Vorinostat bzw. von Rapamycin mit Genen, die durch c-Myc herunterreguliert werden zeigt, dass auch diese Methode zu guten Ergebnissen führt.

Die verschiedenen bioinformatischen Methoden eignen sich für die funktionelle Interpretation der Ergebnislisten. Da die Methoden zu unterschiedlichen Ergebnissen führen, bzw. auf unterschiedliche Schwerpunkte fokussieren, ist es sinnvoll, mehrere Methoden parallel durchzuführen, um eine möglichst effiziente und umfassende Interpretation der Listen zu erzielen.

4 Identifizierung von signifikant angereicherten cis-regulatorischen Motiven in den Promotorsequenzen von ko-regulierten Genen

4.1 Einleitung

Eine der zentralen Fragen der Biologie ist, die Mechanismen, welche die Genexpression regulieren, zu verstehen. Wie und wann werden Gene von welchen Faktoren reguliert? Bei der Genregulation handelt es sich um ein komplexes molekulares Netzwerk, das zahlreiche exakt aufeinander abgestimmte Faktoren beinhaltet. Die spezifische Bindung von Transkriptionsfaktoren an Sequenzmotive im cis-regulatorischen Bereich der DNA ist einer dieser Faktoren. Hierdurch wird die Aktivierung oder Repression der Transkription infolge externer Einflüsse oder entwicklungsspezifischer Stadien ermöglicht. Es gibt zahlreiche experimentelle Untersuchungen zur Interaktion einzelner Transkriptionsfaktoren mit den entsprechenden Bindungsmotiven im Promotorbereich einzelner Gene. Da jedoch ein Transkriptionsfaktor zumeist mehrere Gene reguliert, bzw. ein Gen von mehreren Transkriptionsfaktoren reguliert wird, ist die Untersuchung der Komplexität dieser Interaktionen von großer Bedeutung für die Aufklärung der genregulatorischen Mechanismen. Die vollständige Sequenzierung der Genome verschiedener Organismen und die Entwicklung neuer experimenteller und bioinformatischer Methoden haben die genomweite Identifizierung von Promotorbereichen vorangetrieben und damit die Grundlagen zur computergestützten genomweiten Identifizierung von cis-regulatorischen Sequenzmotiven innerhalb dieser Promotorbereiche ermöglicht.

4.1.1 Genregulation

Mit dem Begriff Genregulation werden alle Prozesse zusammengefasst, die die Steuerung der Genexpression beinhalten. Hiermit wird festgelegt, in welcher Konzentration ein Genprodukt in der Zelle vorliegt und damit an der Umsetzung des Phänotyps beteiligt ist.

Die Genexpression beinhaltet folgende Schritte

- Initiation der Transkription
- Termination der Transkription
- Capping
- Polyadenylierung
- Spleißen
- Transport ins Cytoplasma
- Stabilisierung der mRNA im Cytoplasma
- Initiation der Translation
- Posttranslationale Modifikationen an den synthetisierten Proteinen
- Stabilisierung der Proteins

Die Vielzahl der Schritte ermöglicht eine Regulation auf verschiedenen Ebenen. Die Initiation der Transkription im Kernpromotorbereich ist hierbei einer der wichtigsten Schritte, da hier zunächst die Entscheidung gefällt wird, ob und in welcher Kopienzahl das Gen exprimiert wird.

4.1.2 Promotoren

Als Promotoren werden Sequenzbereiche bezeichnet, die für die Initiation der Transkription eines Gens von Bedeutung sind. Promotoren lassen sich grob in drei Bereiche aufteilen: den Kernpromotor, den proximalen und den distalen Promotor.
Als Kernpromotor wird die genomische Region um die transkriptionelle Startstelle bezeichnet. Es gibt keine genaue Definition über die Länge des Kernpromotors, er wird jedoch im allgemeinen definiert als der DNA-Bereich, der benötigt wird, den Transkriptionsinitiationskomplex zu binden und die Transkription infolge der entsprechenden externen Signale zu starten (Sandelin, Carninci et al. 2007). Elemente, die innerhalb des Kernpromotors auftreten, sind die TATA-Box, das Inr-Element und die CpG-Inseln.

TATA-Box

Die TATA-Box liegt ca. 28-34 bp stromaufwärts von der transkriptionellen Startstelle. An ihre Konsensussequenz TATAA bindet das TATA-Box bindende Protein (TBP), das Teil des Präinitiationskomplexes ist. TATA-Boxen treten häufig in gewebespezifischen Promotoren zusammen mit der Inr-Sequenz auf.

Initiator-Element (Inr)

Das Pyrimidin-reiche Inr-Element wird mit der Konsensussequenz YYANWYY beschrieben. Das hierin enthaltene Adenin liegt im Promotor an Position +1 (Smale and Kadonaga 2003).

Die TATA-Box und das Inr-Element sind die einzigen bisher bekannten Kernpromotor-elemente, die alleine den Präinitiationskomplex rekrutieren und die Transkription initiieren können.

CpG-Inseln

Die CpG-Inseln sind genomische Bereiche, in denen CG-Dinukleotide häufiger auftreten als im restlichen Genom (Gardiner-Garden and Frommer 1987). CpG-Insel assozierte Promotoren treten häufig mit Housekeeping oder ubiquitär exprimierten Genen auf (Schug, Schuller et al. 2005).

Im Kernpromotor binden vorwiegend allgemeine Transkriptionsfaktoren, d.h. Trans-kriptionsfaktoren, die zusammen mit der RNA-Polymerase den Initiationskomplex bilden.

Als distaler bzw. proximaler Promotor werden die Bereiche der DNA beschrieben, die 3’ bzw. 5’ von der Kernpromotorregion liegen. Sie fungieren als Bindungsstellen für spezifische Transkriptionsfaktoren, deren Kombination die zeitliche und räumliche Expression der Gene bestimmt. Die genaue Identifizierung dieser Promotorbereiche gestaltet sich als schwierig, da die Sequenzen verschiedener Promotoren untereinander sehr stark variieren. So hat jedes Gen eine individuelle Zusammensetzung seines Promotors.

4.1.3 Identifizierung von Promotorsequenzen

Innerhalb der letzten Jahre wurden eine Vielzahl bioinformatischer Methoden zur Vorhersage von Promotorbereichen entwickelt. Ein Teil dieser Programme basiert auf der Identifizierung von konservierten Sequenzmotiven wie z.B. der TATA Box, der CAAT-Box, bekannten Transkriptionsfaktor-Bindungsstellen und den CpG-Inseln. (Prestridge 1995; Solovyev and Salamov 1997; Zhang 1998; Knudsen 1999; Ioshikhes and Zhang 2000; Scherf, Klingenhoff et al. 2000; Davuluri, Grosse et al. 2001).

Dieser rein computergestützte Ansatz führt jedoch nur zu einer geringen Anzahl von korrekt vorhergesagten Promotorbereichen, da diese Sequenzmotive nur in einem Teil der Promotoren auftreten. So findet sich z.B. in nur 10-20% der Promotoren eine TATA-Box und in nur 60% der Promotoren CpG-Inseln (Gardiner-Garden and Frommer 1987). Eine TATA-ähnliche Sequenz findet sich dagegen innerhalb des Genoms zufällig alle 250 bp (Bucher 1990).

Andere Programme vergleichen Volllängen-mRNA mit der entsprechenden genomischen DNA und leiten hieraus, basierend auf der Annahme, dass der Promotorbereich unmittelbar vor der Transkriptionsstartstelle liegt, den Promotorbereich ab. Dies ist ein Ansatz, der in vielen Fällen zu einer korrekten Vorhersage des Promotorbereiches führt (Qiu, Ding et al. 2002; Qin, Qiu et al. 2003; Trinklein, Aldred et al. 2003), jedoch durch das Fehlen von vollständigen cDNA-Transkripten limitiert wird. Die Herstellung von cDNA-Banken erfolgt meistens nach der Methode von Gubler und Hoffman (Gubler and Hoffman 1983). Bei dieser Methode werden die mRNA-Transkripte in die komplementären cDNA-Stränge umgeschrieben, ein Vorgang, der bei vielen Transkripten abbricht, bevor das 5'-Ende der zu kopierenden mRNA erreicht ist. Die Ursachen hierfür liegen zumeist in der Ausbildung von Sekundärstrukturen innerhalb des zu kopierenden mRNA-Stranges und einer zu geringen Aktivität der reversen Transkriptase. Hieraus resultieren dann cDNA-Moleküle, denen ein Teil des 5'-Endes der komplementären mRNA fehlt.

Mit neuen experimentellen Ansätzen wie z.B. der *oligo-capping* Methode (siehe Abbildung 16) können inzwischen immer mehr vollständige Transkripte generiert werden, was zu einer verbesserten Vorhersage von Promotorsequenzen geführt hat (Suzuki, Yamashita et al. 2004). Bei der *oligo-capping* Methode wird die cap-Struktur der mRNA am 5´-Ende durch ein synthetisiertes Oligonukleotid ersetzt. Über einen oligo-dT-Primer und einen Primer, der komplementär zu der am 5´-Ende liegenden

Oligonukleotid-Sequenz ist, erfolgt eine Amplifizierung der Sequenzen. So werden nur vollständige mRNA-Transkripte amplifiziert. Die so generierten cDNA-Bibliotheken enthalten einen deutlich geringeren Anteil an unvollständigen cDNA-Sequenzen, der jedoch immer noch bei 30% liegt. In der *DataBase of Transcriptional Start Sites* (DBTSS) finden sich Promotoren, die anhand von cDNA-Sequenzen, die mittels der *oligo-capping* Methode generiert wurden, ermittelt wurden.

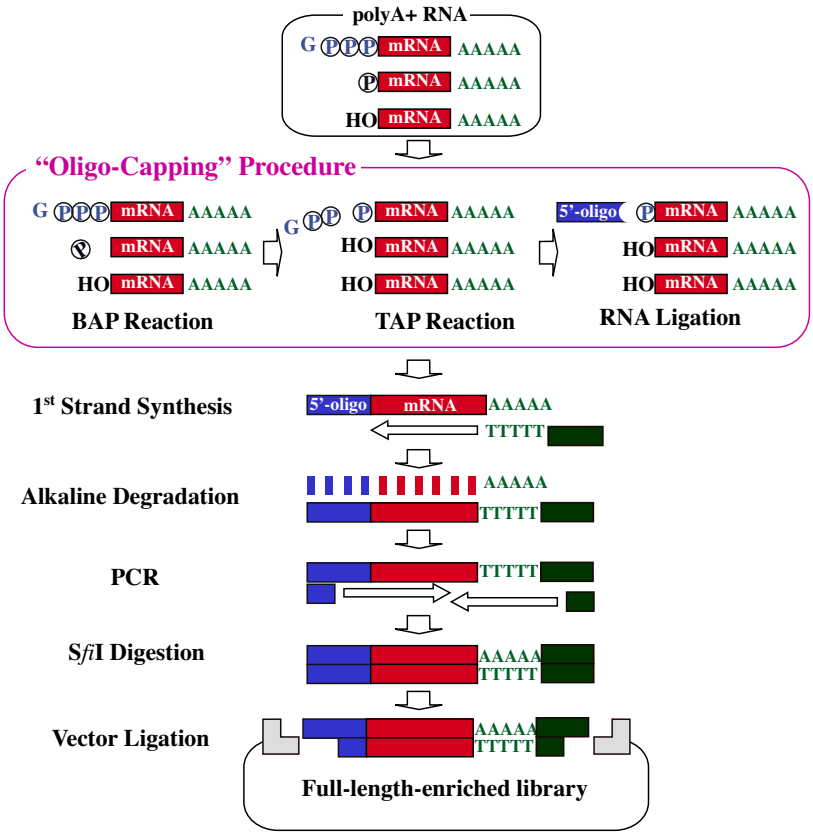


Abb. 16: Herstellung einer cDNA-Klon Bibliothek nach der *oligo-capping* Methode (aus: http://dbtss.hgc.jp/). Im Rahmen der *oligo-capping-procedure* wird die cap-Struktur der mRNA am 5´-Ende durch ein synthetisiertes Oligonukleotid ersetzt. Dies erfolgt in drei Schritte: (1) Um unvollständige mRNA-Moleküle aus den weiteren Reaktionsschritten auszuschließen, werden von den mRNA-Fragmenten bei denen keine cap-Struktur mehr vorliegt, die Phosphatmoleküle am 5´-Ende mittels der bakteriellen alkalischen Phosphatase (BAP) entfernt. (2) Von den mRNA-Molekülen, die eine cap-Struktur haben und somit noch vollständig vorliegen, wird mittels der Gesamt-alkalische Phosphatase (TAP) die cap-Struktur entfernt und anschließend mit der *T4 RNA*-Ligase, welche Phosphate am 5´-Ende als Substrat benötigt, selektiv ein Oligonukleotid am 5´-Ende ligiert. (3) Über einen oligo-dT-Primer und einem Primer, der komplementär zu der am 5´-Ende liegenden Oligonukleotid-Sequenz ist, erfolgt eine Amplifizierung der Sequenzen. So wird gewährleistet, dass nur vollständige mRNA-Sequenzen amplifiziert werden. Diese werden anschließend in einen Vektor ligiert.

Die exakte Bestimmung des Promotorbereiches der Gene ist für die Identifizierung von Transkriptionsfaktor-Bindungsstellen von entscheidender Bedeutung.

4.1.4 Transkriptionsfaktoren und Transkriptionsfaktor-Bindungsstellen

Die Transkriptionsfaktoren spielen eine wichtige Rolle bei der Regulation der Genexpression. Innerhalb der Zelle gibt es eine Vielzahl an Transkriptionsfaktoren. Derzeit sind ca. 1400 humane Transkriptionsfaktoren bekannt, es wird jedoch angenommen, dass die Gesamtzahl der in der Zelle vorliegenden Transkriptionsfaktoren zwischen 1800 und 3000 liegt (Lander, Linton et al. 2001). Als Transkriptionsfaktor werden alle Proteine – außer der RNA-Polymerase – bezeichnet, die zur Initiierung des Transkriptionsvorganges benötigt werden. Sie binden an die Transkriptionsfaktor-Bindungsstellen. Dies sind kurze, ca. 5-15 bp lange, häufig degenerierte Sequenzen, die stromaufwärts oder stromabwärts der Transkriptionsstartstelle oder in dem ersten Intron des Genes, welches sie regulieren, liegen können. Die meisten Transkriptionsfaktor-Bindungsstellen sind jedoch in der proximalen Promotorregion innerhalb von einem Kilobasenpaar stromaufwärts der Transkriptionsstartstelle zu finden (Zhang 1998). Die Funktion der Transkriptionsfaktoren ist oft unabhängig von der Orientierung, mit der sie an die DNA binden. Transkriptionsfaktoren bzw. ihre Komplexe agieren entweder direkt oder durch Mediatoren mit der RNA-Polymerase im Initiationskomplex. Bei den Transkriptionsfaktoren können allgemeine und spezifische Faktoren unterschieden werden.

Als allgemeine Transkriptionsfaktoren werden diejenigen bezeichnet, die zusammen mit der RNA-Polymerase den Initiationskomplex bilden und mit diesem am Kernpromotor binden.

Die spezifischen Transkriptionsfaktoren, welche in distalen und proximalen Promotorbereichen binden, werden noch einmal in zwei weitere Gruppen unterteilt, die konstitutiven und die induzierbaren Transkriptionsfaktoren.

Die konstitutiven Transkriptionsfaktoren erkennen kurze Konsensus-Elemente, die sich "oberhalb" oder im 5’-Bereich des Transkriptionsstarts befinden (z.B. Sp1, das an die GC Box bindet). Diese Faktoren sind ubiquitär und wirken auf jeden Promotor, der eine entsprechende Bindungsstelle besitzt. Sie erhöhen die Effizienz der Initiation.

Die induzierbaren Faktoren funktionieren ähnlich wie die konstitutiven Faktoren, haben jedoch eine regulative Rolle. Sie werden zeit- und gewebeabhängig synthetisiert oder aktiviert. Die von ihnen gebundenen Sequenzen nennt man *response elements*.

Die Unterteilung in konstitutive und induzierbare Transkriptionsfaktoren ist nicht strikt, da es Transkriptionsfaktoren gibt, die beide Funktionen übernehmen können.
Die Struktur und Wirkungsweise der Transkriptionsfaktoren und der Transkriptionsfaktor-Bindungsstellen beinhalten eine Reihe von Regulationsmechanismen für die Genexpression. So kann durch die Kombination verschiedener induzierbarer Transkriptionsfaktoren die Genexpression nach erfolgtem externem Stimulus gezielt räumlich und zeitlich gesteuert werden. Auch die Kombination von mehreren Transkriptionsfaktoren in einem Komplex erhöht die Flexibilität der Regulation. So wird durch die Komplexbildung die gesamte Kontaktfläche der Bindung an die DNA erhöht und dadurch die Sequenzspezifität jeder individuellen Bindung gelockert. Dies ermöglicht dem Komplex ein breiteres Spektrum an Transkriptionsfaktor-Bindungsstellen zu erkennen (Chen 1999). Durch die degenerierten Bindungssequenzen von Transkriptionsfaktoren können unterschiedlich hohe Bindungsaffinitäten erzielt werden, so dass verschiedene Gene durch den gleichen Transkriptionsfaktor unterschiedlich stark transkribiert werden können (Stormo 2000).

Die Eigenschaft der Transkriptionsfaktor-Bindungsstellen, sehr kurz und degeneriert zu sein, erschwert jedoch die zuverlässige bioinformatische Erkennung innerhalb der Promotorsequenz, da die Chance, dass ein bestimmtes Sequenzmuster auch zufällig – und damit ohne funktionelle Bedeutung – gefunden wird, deutlich erhöht ist.

Experimentelle Methoden zur Identifizierung von Transkriptionsfaktor-Bindestellen

Es gibt eine Reihe von experimentellen Methoden zur Untersuchung der Wechselwirkung zwischen Transkriptionsfaktoren und ihren DNA-Bindestellen und damit der Identifizierung von Transkriptionsfaktor-Bindungsstellen. Hierzu gehören z.B. die Gel-Shift-Analyse und das DNase I Schutz-Experiment (Galas und Schmidt 1978). Hochdurchsatz-Methoden zur Identifizierung von Transkriptionsfaktor-Bindungsstellen wie z.B. die *ChIP-on-chip* Methode erlauben zwar, Transkriptionsfaktor-Bindungsstellen für große Teile des Genoms zu bestimmen, allerdings kann dabei die Lokalisation der Bindungsstellen nur auf Fragmente mit einer Länge von ca. 200 bp eingegrenzt werden (Ren, Robert et al. 2000; Iyer, Horak et al. 2001; Lieb, Liu et al. 2001).
Eine viel versprechende neuere Methode erscheint die *'paired-end ditag'* (PET) Strategie, die genomweit sowohl im proximalen und distalen Promotorbereich als auch

räumlich von der Transkriptionsstartstelle entfernt liegende Transkriptionsfaktor-bindungsstellen identifizieren kann (Wei, Wu et al. 2006).

Bioinformatische Methoden zur Identifizierung von Transkriptionsfaktor-Bindestellen

Neben den experimentellen Ansätzen ist die bioinformatische Vorhersage der Bindungsstellen von Bedeutung. Da funktionell relevante Transkriptionsfaktor-bindungsstellen aufgrund ihrer Sequenzeigenschaften sehr schwierig zu detektieren sind, ist das Ziel hierbei, innerhalb einer vorgegebenen Nukleotidsequenz möglichst viele funktionelle Transkriptionsfaktor-Bindungsstellen zu erkennen, und dabei gleichzeitig möglichst wenig falsche Vorhersagen zu machen. Bereits bekannte Bindungsstellen dienen als Basis für die Erstellung eines Modells, welches die degenerierte Bindungsstelle möglichst genau beschreiben soll.

Pos	A	C	G	T	IUPAC
01	6	2	8	1	R
02	3	5	9	0	S
03	0	0	0	17	T
04	0	0	17	0	G
05	17	0	0	0	A
06	0	16	0	1	C
07	3	2	3	9	T
08	4	7	2	4	N
09	9	6	1	1	M
10	4	3	7	3	N
11	6	3	1	7	W

code	description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

Abb. 17: Matrix zum Erstellen der Konsensus-Sequenz des Transkriptionsfaktor AP-1 im IUPAC Code. Auf der linken Seite ist die Matrix zur Erstellung der Konsensus Sequenz des Transkriptionsfaktors AP-1. Basierend auf experimentel nachgewiesenen Bindungsstellen wird für jede Position gezählt, wie häufig jede der 4 Basen an dieser Position auftritt. Diese Auflistung wird dann in den IUPAC-Code übertragen. (aus Wikipedia: http://en.wikipedia.org/wiki/Sequence_motif#Matrices) rechts: IUPAC Code

Konsensus-Sequenz

Ein Ansatz zur Beschreibung von Bindungsstellen ist die Bildung einer Konsensus-Sequenz. Dazu werden die bekannten, experimentell nachgewiesenen Bindungsstellen miteinander verglichen und für jede Position beschrieben, welche Base am häufigsten auftritt. Bei diesem Ansatz geht jedoch ein wichtiger Teil an Information verloren, da die Variabilität der einzelnen Positionen nicht erfasst wird (Roulet, Fisch et al. 1998). Ein Teil dieser Information kann durch die Verwendung des IUPAC Codes erhalten bleiben, da dieser für jede Position das Auftreten unterschiedlicher Nukleotide beschreiben kann. In Abbildung 17 wird der IUPAC Code des Transkriptionsfaktors AP-1 dargestellt. Hiermit wird jedoch immer noch nicht beschrieben, mit welcher Häufigkeit andere Nukleotide an den einzelnen Positionen auftreten können. Mit der Verwendung der Konsensussequenz und des IUPAC Codes zur Identifizierung von Transkriptionsfaktor-Bindungsstellen werden zwar keine falsch positiven Sequenzen gefunden, d.h. es treten nur Sequenzen auf, die auch schon in die Erstellung der Konsensussequenz mit eingegangen sind, aber einige Motive werden nicht gefunden, da sie Nukleotide beinhalten, die durch die Konsensussequenz nicht berücksichtigt werden.

***Positions-specific scoring matrices* (PSSMs)/*Position weight matrices* (PWM)**

Eine weitere Möglichkeit der Signalerkennung bieten die Positions-spezifischen *scoring matrices* (PSSMs). Die PSSMs zur Sequenzanalyse wurden erstmals von Stormo, Schneider et al. (1982) zur Untersuchung der Transkriptionsstartstellen in *Escherichia coli* verwendet. Auch für diese Methode werden experimentell nachgewiesene Bindungsstellen als Grundlage verwendet. Für jede Transkriptionsfaktor-Bindungsstelle wird eine PSSM erstellt, aus der ein Wert ermittelt wird, der die Wahrscheinlichkeit angibt, mit der ein Nukleotid an einer bestimmten Stelle innerhalb der Bindungssequenz auftritt. Um diesen Wert zu berechnen, wird an jeder Stelle der Sequenz die Anzahl der verschiedenen Nukleotide, die in den vorgegebenen Bindungssequenzen auftreten, aufsummiert und als Prozentwert in die PSSM übertragen (s. Abbildung 18).
Anhand von PSSM lässt sich eine Bewertung einzelner Sequenzabschnitte vornehmen, hinsichtlich der Wahrscheinlichkeit, dass eine bestimmte Bindungsstelle vorliegt. In der Praxis wird die PSSM-Matrix meistens in eine *position weight matrice* (PWM) konvertiert. Die Konvertierung erfolgt, indem die normalisierten Häufigkeiten auf eine

Log-Skala transformiert werden. Ein quantitativer Wert zur Bewertung der potentiell relevanten Bindungsstelle wird ermittelt, indem die PWM Werte für die vorhandenen Nukleotide summiert werden.

Position in der Sequenz		1	2	3	4	5	6
Experimentell bestimmte TFBS		A	C	A	T	G	C
		A	T	A	C	G	C
		A	A	T	T	G	A
		A	T	A	T	C	C
		A	C	A	C	G	A
		A	G	A	T	C	C
		A	T	T	C	G	C
		A	T	T	T	C	C
PSSM	A	8	1	5	0	0	2
	C	0	2	0	3	3	6
	T	0	4	3	5	0	0
	G	0	1	0	0	5	0

Abb. 18: Beispiel für die Erstellung einer PSSM. Die experimentell ermittelten Bindungsstellen werden untereinander aufgelistet. Für jede Position der Sequenz wird die Anzahl der verschiedenen Nukleotide aufsummiert und in die PSSM übertragen

Da PSSM/PWM einen größeren Anteil der Sequenzinformation nutzen als Konsenussequenzen, können sie potentielle Transkriptionsfaktor-Bindungsstellen genauer vorhersagen. Die Qualität von PSSM/PWM und damit ihre „Fähigkeit“, potentielle Transkriptionsfaktor-Bindungsstellen zu finden, ist jedoch abhängig von der Genauigkeit des Modells, welches zur Beschreibung der PWM/PSSM vorliegt. Dies ist wiederum abhängig von der Anzahl der zugrunde liegenden experimentell bestimmten Bindungsstellen (Roulet, Fisch et al. 1998). Da für viele Transkriptionsfaktor-Bindungsstellen die Modelle nur anhand einer geringen Anzahl von experimentell bestimmten Bindungsstellen erstellt werden konnten, führen auch PSSM/PWM häufig zu falsch positiven Vorhersagen (Jolly, Chin et al. 2005; Tompa, Li et al. 2005).

Es gibt eine Reihe von computergestützten Programmen, welche Transkriptionsfaktor-Bindungsstellen innerhalb von Sequenzen anhand von PSSM/PWM identifizieren z.B.

TESS (http://www.cbil.upenn.edu/tess/), MATCH™ (Kel, Gossling et al. 2003), MatInspector (Quandt, Frech et al. 1995), MATRIX SEARCH (Chen, Hertz et al. 1995), SIGNAL SCAN (Prestridge 1996) und rVISTA (Loots, Ovcharenko et al. 2002).

4.1.5 Phylogenetisches *Footprinting*

Das phylogenetische *Footprinting* hat sich als eine bedeutende Methode im Rahmen der bioinformatischen Identifizierung von Transkriptionsfaktor-Bindungsstellen etabliert. Zur besseren Fokussierung auf funktionell relevante Transkriptionsfaktoren und der Reduktion der Zahl falsch positiver Bindungsstellen wird hierbei auf orthologe nicht-kodierende Sequenzbereiche fokussiert, die zwischen verschiedenen Spezies konserviert sind. Orthologe Sequenzen sind Sequenzen, die sich bei verschiedenen Arten im Rahmen der Evolution aus einer gemeinsamen Vorläufersequenz entwickelt haben und von denen angenommen wird, dass sie ihre Funktion behalten haben (Tagle, Koop et al. 1988).

Das phylogenetische *Footprinting* basiert auf zwei Annahmen:

1. Die Funktionen und DNA Bindungseigenschaften von Transkriptionsfaktoren sind zwischen verschiedenen Spezies gut konserviert
2. Nicht-kodierende DNA-Sequenzen, die für die Regulation der Genexpression notwendig sind, unterliegen einem höheren evolutionären Selektionsdruck als Sequenzbereiche ohne funktionelle Relevanz. Somit beinhalten diese Sequenzbereiche, zu denen auch die Transkriptionsfaktor-Bindungsstellen gehören, eine geringere Rate an Mutationen (Neph and Tompa 2006).

Die Bedeutung von konservierten nicht-kodierenden Bereichen für die Genregulation wurde von Wasserman and Fickett (1998) demonstriert. Sie zeigen, dass 98% der experimentell bestätigten humanen Skelettmuskel-spezifischen Transkriptionsfaktor-Bindungsstellen in nur 19% der nicht-kodierenden Sequenzbereichen liegen, die einen sehr hohen Konservierungsgrad mit dem Mausgenom aufweisen.
Auch zeigt der Vergleich der Konservierungsraten von Maus- und Menschgenom, dass der Anteil an kurzen konservierten Segmenten wesentlich höher ist als durch die für Protein kodierenden Sequenzen alleine erklärt werden könnte (Waterston, Lindblad-Toh

et al. 2002). 2/3 dieser Sequenzen liegen in nicht-kodierenden Bereichen. Diese konservierten nicht-kodierenden Bereiche treten weitgehend einmalig im Genom auf, d.h. sie sind nicht repetitiv.

Die Auswahl der Organismen, welche für das phylogenetische *Footprinting* herangezogen werden sollte, ist abhängig von der Fragestellung. So ist z.B. bei Mammalia und Knochenfischen, die mit 450 mya (*m*illion *y*ears *a*go) sehr weit divergieren, der Anteil der orthologen Gene, die konservierte Elemente in den nicht-kodierenden Regionen haben, nur sehr gering. Der Vergleich von weit divergierenden Organismen eignet sich dagegen gut für die Untersuchung von Genen aus der frühen Embryonalentwicklung. Ein Vergleich von Organismen, die sehr nah verwandt sind, wie z.B. Mensch und Schimpanse, zeigen dagegen so geringe Unterschiede, dass hiervon kaum profitiert werden kann (Lenhard, Sandelin et al. 2003). Zur Identifizierung funktionell relevanter Bereiche im humanen Genom wird häufig der Vergleich Mensch und Maus verwendet, die eine Diversität von 75 mya haben. Ungefähr 40% des gesamten Mausgenoms kann mit dem humanen Genom verglichen werden und 80% der Mausgene haben ein identifizierbares orthologes Gen im humanen Genom (Waterston, Lindblad-Toh et al. 2002). Somit ist die Divergenz zwischen beiden Genomen klein genug, um noch orthologe Bereiche zu finden, aber doch groß genug, um funktionelle von nicht-funktionellen Bereichen zu unterscheiden. Es gibt bereits einige Arbeiten, die mit dem Mensch-Maus Vergleich funktionell relevante Sequenzmotive gefunden haben (Hardison, Slightom et al. 1997; Loots, Locksley et al. 2000; Dermitzakis and Clark 2002; Elnitski 2003).

Das phylogenetische *Footprinting* beinhaltet 3 Schritte:

1. Auswahl geeigneter orthologer Sequenzen
2. Vergleichen der orthologen Promotorsequenzen
3. Identifizieren von signifikant konservierten Motiven

1. Auswahl geeigneter orthologer Sequenzen

Da Transkriptionsfaktor-Bindungsstellen am häufigsten im Sequenzbereich um die Transkriptionsstartstelle auftreten, werden vorwiegend die Promotorsequenzen zur Suche nach funktionellen Transkriptionsfaktor-Bindungsstellen verwendet. Die

Probleme und Strategien bei der Identifizierung von Promotorsequenzen wurden bereits in Kapitel 4.1.3 beschrieben.

2. Vergleichen der orthologen Promotorsequenzen

Ein wichtiger Schritt des phylogenetischen *Footprinting*s ist der Vergleich der orthologen Sequenzen. Bei dem Sequenz-Alignment werden 2 oder mehr biologische Sequenzen auf ihre Ähnlichkeit hin untersucht. Hierbei unterscheidet man das lokale und das globale Alignment.

Lokales Alignment

Bei dem lokalen Alignment wird nach möglichst optimalen Übereinstimmungen in Teilbereichen der Sequenz gesucht, es wird jedoch keine Übereinstimmung auf der gesamten Länge der Sequenz berücksichtigt. Dies ermöglicht, innerhalb von längeren Sequenzen kürzere konservierte Elemente zu finden.

Globales Alignment

Das globale Alignment sucht nach möglichst optimaler Übereinstimmung auf der gesamten Länge der zu vergleichenden Sequenzen. Globale Alignments werden hauptsächlich verwendet, wenn die zu untersuchenden Sequenzen ähnlich lang sind und starke Sequenzhomologien erwartet werden. Die Verwendung des globalen Alignments basiert auf der Annahme, dass wichtige funktionelle Sequenzen im Laufe der Evolution kollinear geblieben sind, d.h. die gleiche Ausrichtung und Anordnung haben. Eine Annahme, die häufig nicht zutrifft, da die Anordnung von Sequenzbereichen häufig im Laufe der Evolution durch Translokation oder Inversion verändert worden ist (Sankoff and Cedergren 1973).

3. Identifizieren von signifikant konservierten Motiven

Innerhalb der konservierten Sequenz wird im letzten Schritt für eine Gruppe von Genen nach Sequenzmotiven gesucht, die signifikant im Vergleich zu einem Hintergrund-Genset angereichert sind.

Im Folgenden wird die Etablierung einer Methode zur Identifizierung von cis-regulativen Motiven beschrieben.

Hierzu wurden folgende Einzelschritte etabliert

- Erstellung einer Datenbank mit orthologen Promotorsequenzen von humanen bzw. Mauspromotoren
- Maskierung von repetitiven Sequenzelementen
- Alignment der orthologen Sequenzen mit dem Programm DiAlign2
- Untersuchung der konservierten humanen Sequenzbereiche auf potentielle Transkriptionsfaktor-Bindungsstellen
- Korrektur der Anzahl der Bindungsstellen auf die Länge der konservierten Promotorsequenzen
- Untersuchung von einer Gruppe von ko-regulierten Genen auf die Anreicherung von Transkriptionsfaktor-Bindungsstellen im Vergleich zu einem Hintergrundset

4.2 Programme und Datenbanken

4.2.1 Repeatmasker

Repeatmasker (www.repeatmasker.org) untersucht DNA-Sequenzen auf bekannte repetitive Sequenzbereiche (z.B. GC- und Alu-Motive) und Regionen geringer Komplexität (z.B. lange GGGGGG-Abschnitte). Als Ergebnis erhält der Nutzer eine Liste der Positionen der einzelnen Wiederholungen und eine modifizierte Version der eingegebenen Sequenz, in der die Wiederholungssequenzen „maskiert“ sind, d.h. sie sind durch N´s ersetzt.

4.2.2 DiAlign-2.2

Dialign 2.2 ist ein Programm zum Alignment von mehreren DNA- und Protein-sequenzen. Das Programm, dessen Name für „*Diagonal Alignment*“ steht, kombiniert lokale und globale Alignment-Eigenschaften, wodurch Sequenzen korrekt verglichen werden, bei denen lokale Ähnlichkeiten räumlich durch unverwandte Sequenzen getrennt sind. Dies ist eine Situation, die häufig auch für konservierte Transkriptions-faktorbindungsstellen (TFBS) in Promotorsequenzen gilt (Werner, Fessele et al. 2003).

Der Algorithmus läuft in vier Phasen ab:

1. Bildung von Fragmenten durch paarweises Alignment

Zuerst werden Regionen mit starker lokaler Homologie in allen Sequenzenpaaren gesucht. Solche homologen Regionen werden als Fragmente oder Diagonalen bezeichnet (s.Abbildung 19).

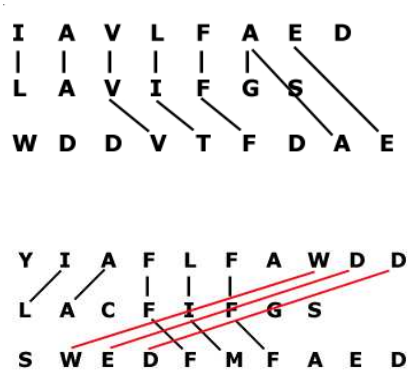


Abb. 19: DiAlign-2.2. Beispiel für das Auffinden von homologen Sequenzpaaren bei dem Alignment von Sequenzen verschiedener Organismen

2. Gewichtung der Ähnlichkeiten

Jeder möglichen Diagonale wird ein Gewicht zugeordnet, das der Ähnlichkeit der Regionen entspricht.

3. Fragmente nach ihren Ähnlichkeitsbewertungen sortieren

4. Konsistenzprüfung

Das Programm versucht anschließend ein multiples Alignment aufzubauen, das dann aus einer konsistenten Ansammlung solcher Diagonalen mit maximalem Gewicht besteht.

(Morgenstern 1999)

4.2.3 MATCH™

Das Programm MATCH™ sucht innerhalb von DNA-Sequenzen nach potentiellen Transkriptionsfaktor-Bindungsmotiven. Hierbei verwendet das Programm *position weight matrices* (PWM), durch die die Bindungsstellen beschrieben werden. Die Sequenz wird über die Benutzeroberfläche des web-basierten Programms eingegeben.

Die folgende Abbildung zeigt die Benutzeroberfläche des Programms.

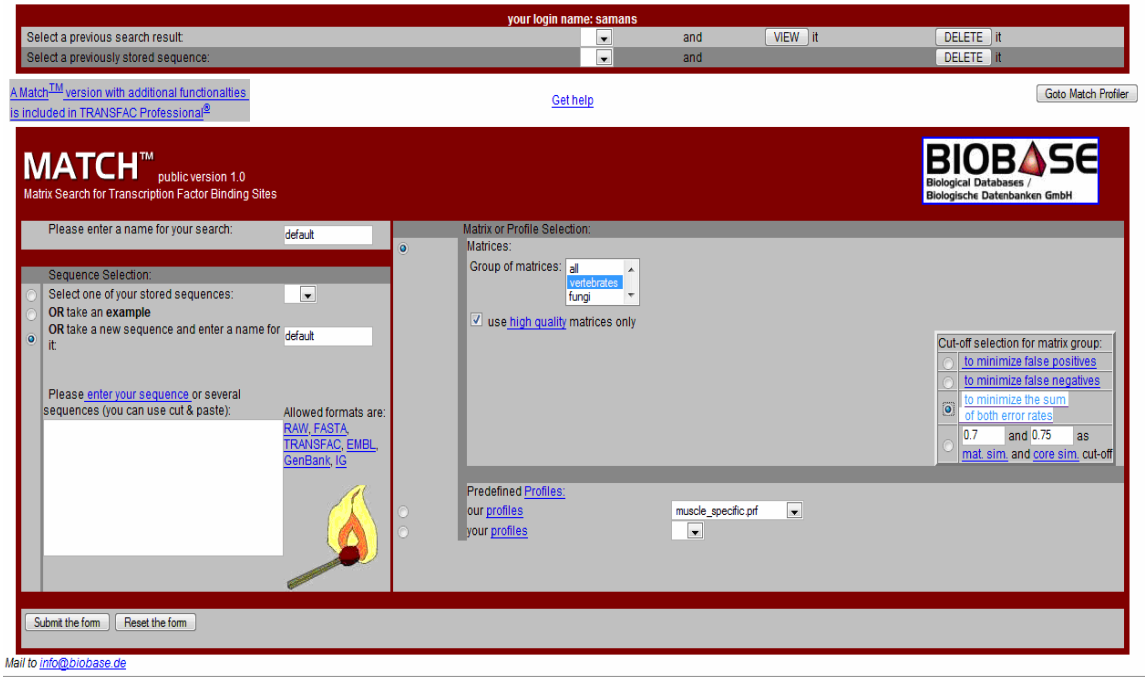


Abb. 20: Benutzeroberfläche der Software MATCHTM. Die Software MATCHTM wurde zur Suche von TFBS auf Basis von PWM verwendet. MATCHTM verwendet hierbei die Datenbank TRANSFAC®, welche derzeit die größte Sammlung von PWM zur Verfügung stellt. (http://www.gene-regulation.com/cgi-bin/pub/programs/ match/bin/match.cgi?)

Die Identifikation von Sequenzbereichen als potentielle Bindungsmotive erfolgt über zwei verschiedene *scores*, dem *matrix similarity score* und dem *core similarity score*. Beide Parameter werden vom Benutzer vorgegeben, der damit die Stringenz der Suche festlegt. Der *matrix similarity score* beschreibt die Qualität der Übereinstimmung der Sequenz mit der gesamten PWM, der *core similarity score* ist ein Maß für die Übereinstimmung des *core* Bereiches der PWM (die 5 am höchsten konservierten Nukleotide).

Die Berechnung der beiden Scores funktioniert folgendermaßen:

Für jede PWM werden alle möglichen Übereinstimmungen der zugehörigen *core*-Matrix mit der zu untersuchenden Sequenz berechnet.

Für jede dieser Untersequenzen wird die Startposition in der Sequenz und der *core similarity* Wert in einer Tabelle erfasst.

Bei den *core similarity* scores, die über dem vom Benutzer gewählten Schwellenwert liegen, wird die Untersequenz an beiden Seiten verlängert, so dass sie der Matrixlänge entspricht. Dann wird der *matrix similarity score* berechnet. Sequenzbereiche, die hier auch über dem Schwellenwert des *matrix similarity scores* liegen, werden in die Ausgabetabelle mit aufgenommen. Beide Scores gehen von 0-1, wobei 1 eine exakte Übereinstimmung zwischen Sequenz und PWM ist.

MATCH™ stellt für alle PWMs der TRANSFAC®-Datenbank vordefinierte Schwellenwerte zur Verfügung, die daraufhin optimiert wurden, die Anzahl der falsch positiven bzw. falsch negativen Bindungsstellen zur reduzieren. Die Optimierung erfolgte, indem die Auffindungssrate in Exon 2, ein Bereich, in dem keine TFBS auftreten sollten, mit unterschiedlichen Schwellenwerten untersucht wurde.

Das Ergebnis der Auswertung wird im HTML-Format erstellt. Ein Beispiel hierfür ist in Abbildung 21 dargestellt.

Scanning sequence ID: Hs_100058;
View a graphic of the following search results

matrix identifier	position (strand)	core match	matrix match	sequence (always the (+)-strand is shown)	factor name
V$LMO2COM_02	26 (-)	0.973	0.975	cccTATCCg	Lmo2 complex
V$CP2_01	56 (-)	1.000	0.937	CTGGGtcctgc	CP2
I$TTK69_01	59 (+)	1.000	1.000	ggTCCTGc	Ttk 69K
F$STUAP_01	104 (+)	0.969	0.973	caCCGCGtcc	StuAp
V$HNF4_01	227 (-)	1.000	0.800	gagccggCTTTGgccaagg	HNF-4
V$HNF4_01	234 (+)	0.848	0.817	ctttggcCAAGGtgccttc	HNF-4
V$IK1_01	263 (-)	0.837	0.947	gtggTTCCCagag	Ik-1
V$RFX1_02	264 (-)	0.982	0.948	tgGTTCCcagagcgtcgc	RFX1
V$CAAT_01	292 (-)	0.950	0.947	tgACTGGctcnn	CCAAT box
V$HNF4_01	307 (-)	0.848	0.786	cccggggCCTTGggcgcct	HNF-4
F$PACC_01	351 (-)	0.976	0.949	ggcggggcCTGGCagga	PacC
V$CETS1P54_01	362 (+)	0.974	0.947	gCAGGAcgtg	c-Ets-1(p54)
V$HNF4_01	460 (+)	0.752	0.806	cggggtgCAGAGatcagac	HNF-4
V$GATA3_03	468 (+)	0.977	0.962	agaGATCAga	GATA-3
I$HAIRY_01	569 (+)	1.000	0.963	cacgCACGCgtccc	Hairy
V$USF_Q6	571 (+)	0.960	0.922	cGCACGcgtc	USF
V$PAX4_01	581 (+)	0.977	0.812	ccggcTCACGcgtccnnnnnc	Pax-4
I$HAIRY_01	583 (+)	1.000	0.939	ggctCACGCgtccn	Hairy
V$USF_Q6	585 (+)	1.000	0.953	cTCACGcgtc	USF
V$CP2_01	611 (+)	0.987	0.881	gcccctACCAG	CP2
V$CETS1P54_01	636 (+)	0.974	0.956	nCAGGAtgtc	c-Ets-1(p54)
I$ELF1_01	730 (+)	0.949	0.906	cctcctGGTCTtgtgc	Elf-1

Abb. 21: Ausgabeformat der webbasierten Software MATCH™. Die Abbildung stellt die Ergebnisse der Software MATCH™ für die Promotorsequenz eines Gens dar. Jede Reihe der Tabelle stellt eine potentielle TFBS dar. Für jede Bindungsstelle wird tabellarisch der Name der Matrix, die Position innerhalb der eingegebenen Sequenz, der Strang auf dem die Sequenz liegt, der *core similarity score* und der *matrix similarity score*, die Sequenz und der Name des Transkriptionsfaktors ausgegeben.

4.2.4 R/Bioconductor

Das Skriptsprache R bzw. das *Bioconductor*-Projekt wird in Kapitel 2.1.1 beschrieben.

4.2.5 Perl

Perl ist eine freie, plattformunabhängige Programmiersprache. Sie wurde ursprünglich als Werkzeug zur Verarbeitung und Manipulation von Textdateien insbesondere bei System- und Netzwerkadministration verwendet (z.B. Auswertung von Logdateien). Inzwischen ist Perl die in der Bioinformatik am häufigsten verwendete Programmiersprache, u.a. wegen ihrer Stärken bei der String-Behandlung und Textbearbeitung und beim Verbinden von Prozessen.

4.2.6 TRANSFAC®

Die derzeit größte Sammlung an PWMs zur Beschreibung von TFBS ist die TRANSFAC®-Datenbank. In ihr sind Transkriptionsfaktoren und ihre bisher bekannten Bindungsstellen bzw. die daraus abgeleiteten PWMs enthalten. Verschiedene Programme, z.B. auch MATCH™, greifen bei der Suche nach potentiellen TFBS auf diese Datenbank zurück.

4.2.7 Cold Spring Harbor Laboratory Mammalian Promotor Database (CSHLmpd)

Die *Cold Spring Harbor Laboratory Mammalian Promotor Database* (CSHLmpd) ist eine Internetdatenbank, die Promotorsequenzen von humanen, Maus- und Ratten-Genen enthält. Tabelle 6 gibt einen Überblick über die Anzahl der derzeit in der Datenbank vorliegenden Promotoren.

Tab. 6: Überblick über die Anzahl der in der CSHLmpd Datenbank vorliegenden Promotorsequenzen
(aus Yamashita, Suzuki et al. (2006))

	No. of genes/ no. of RefSeq	No. of promoters	No. of TSSs	No. of clones
Human	15262/19753	30964	452117	1359000
Mouse	14162/14746	19023	149876	364487
Zebrafish	3061/3075	3382	15198	32263
Malaria	1527/NA	NA	6906	10236
Schyzon	3635/NA	NA	14029	22923

Die Datenbank wurde erstellt, indem zunächst experimentell bestimmte Promotorsequenzen aus den Datenbanken EPD (*Eukaryotic Promoter Database*) und DBTSS (*Database of Transcriptional Start Site*) bzw. durch Sequenzanalyse bestimmte Promotorsequenzen aus der Datenbank *Genbank*, zusammengestellt wurden. Dazu wurden weitere Promotorsequenzen mittels des Programms *FirstEF* vorhergesagt. In einem komplexen Pipelineverfahren, welches in Abbildung 22 dargestellt ist, wurden diese Promotorsequenzen dann weiter auf ihre Genauigkeit (der Lokalisation) untersucht. Der genaue Prozess wird in Xuan, Zhao et al. (2005) beschrieben.

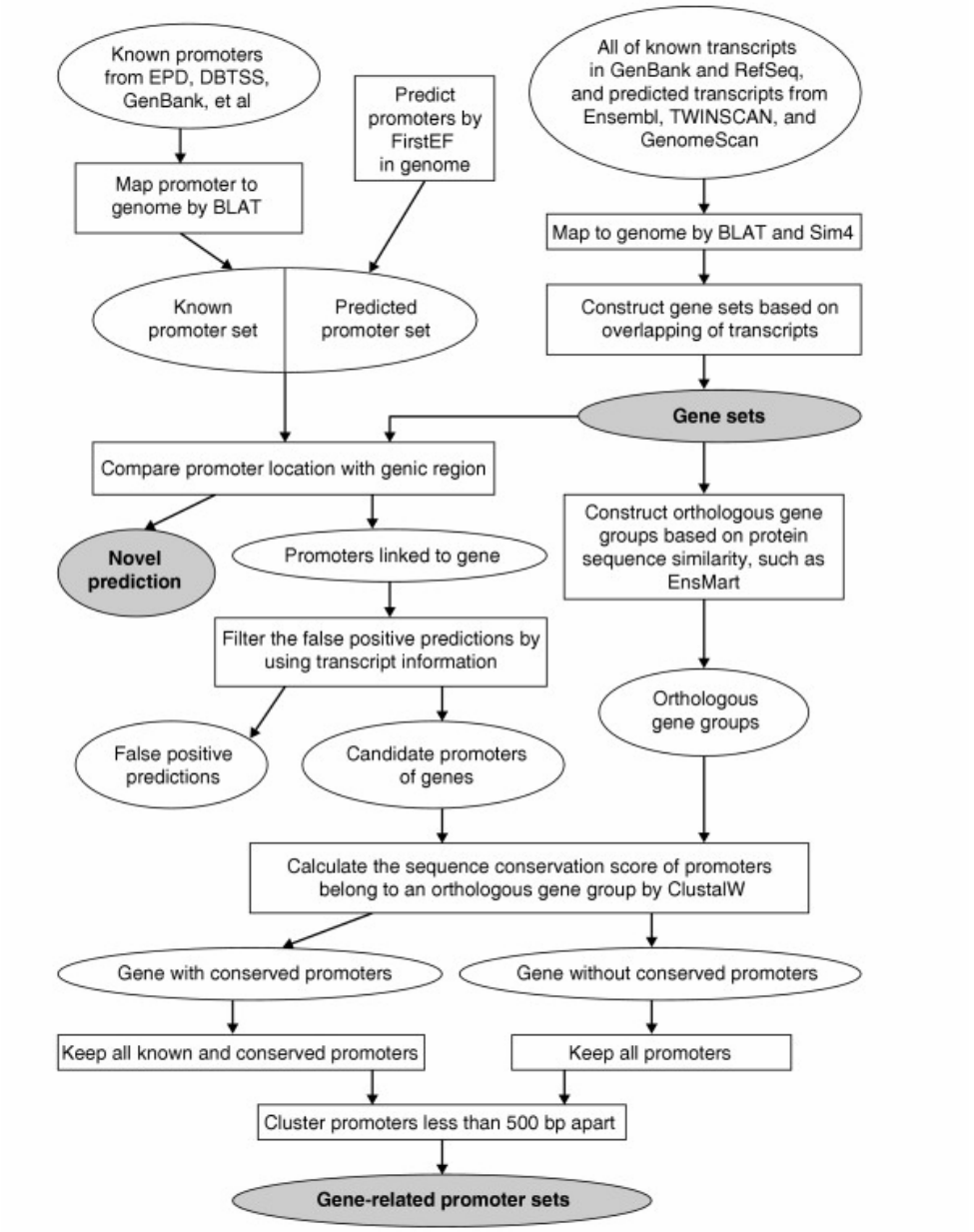


Abb. 22: Pipeline zur Identifizierung von Promotorsequenzen, die in die Datenbank DBTSS integriert wurden. Das genaue Verfahren wird in Xuan, Zhao et al. (2005) beschrieben.

4.3 Etablierung der Methode

4.3.1 Erstellung einer Tabelle orthologer Promotorsequenzen von Mensch- und Mauspromotoren

Als Basis für das phylogenetische *Footprinting* werden Mensch- und Mauspromotoren gewählt. Diese werden aus der CSHLmpd Datenbank heruntergeladen. Die Auswahl der Sequenzen orientiert sich an den Genen des humanen bzw. des Maus-cDNA-Mikroarrays, die am Institut für Molekularbiologie und Tumorforschung (IMT) hergestellt werden, da die ersten Auswertungen, mit denen diese Methode getestet werden soll, auf Versuchen basieren, die mit diesen zwei Mikroarray-Plattformen durchgeführt wurden. Insgesamt wurden 8964 orthologe Promotorpaare gefunden. Bei Sequenzen mit alternativen Promotoren werden die in der Datenbank als "*best*" definierten Sequenzen gewählt. Als Sequenzbereich wird -700 bis +300 Basenpaare um die transkriptionelle Startstelle (TSS) gewählt.

4.3.2 Maskierung von repetitiven Sequenzelementen

Repetitive Sequenzen innerhalb der Promotorsequenzen werden mit dem Programm *Repeatmasker* maskiert. Dies ist notwendig, da repetitive Sequenzen häufig in Promotorsequenzen auftreten und zu einem hohen Grad an Konservierung führen würden, die jedoch keine funktionelle Bedeutung hinsichtlich der Suche nach TFBS hätten. Die Anwendung von *Repeatmasker* erfolgt auf der Internetseite des Programms mit den vorgegebenen Standardeinstellungen. Die Sequenzen werden im Fasta-Format eingegeben. Als Ergebnis erhält man die Sequenzen, ebenfalls im Fasta-Format, in denen die repetitiven Bereiche durch N´s maskiert sind.

4.3.3 Alignment der orthologen Sequenzen

Die orthologen Promotorsequenzen werden mit dem Programm *DiAlign-2.2* auf konservierte Sequenzbereiche untersucht. Dieses Programm wird gewählt, da es sowohl ein lokales als auch ein globales Alignment durchführt. Somit kann es lokale Ähnlichkeiten, die räumlich durch unverwandte Sequenzen getrennt sind, aber auch Sequenzbereiche, in denen funktionelle Elemente ihre Positionen getauscht haben erkennen. Das

Programm wird hierzu lokal auf einer UNIX-Plattform installiert, wodurch eine automatisierte Analyse der 8964 orthologen Promotorpaare ermöglicht wird.
Die jeweiligen orthologen Sequenzen werden im Fasta-format in einzelnen Dateien abgespeichert. Mittels des Befehls ~/dialign2_dir/dialign2-2 <Dateiname> wird das Alignment gestartet. Für jedes Alignment wird eine Datei ausgegeben. In dieser Datei werden konservierte Sequenzbereiche als Großbuchstaben dargestellt und nicht-konservierte Bereiche in Kleinbuchstaben. Ein Beispiel für die Ausgabedatei ist in Abschnitt 10.4 im Anhang dargestellt. Der mittlere Konservierungsgrad liegt bei 69,5%. Die Verteilung der konservierten Sequenzlängen ist in Abbildung 23 dargestellt.

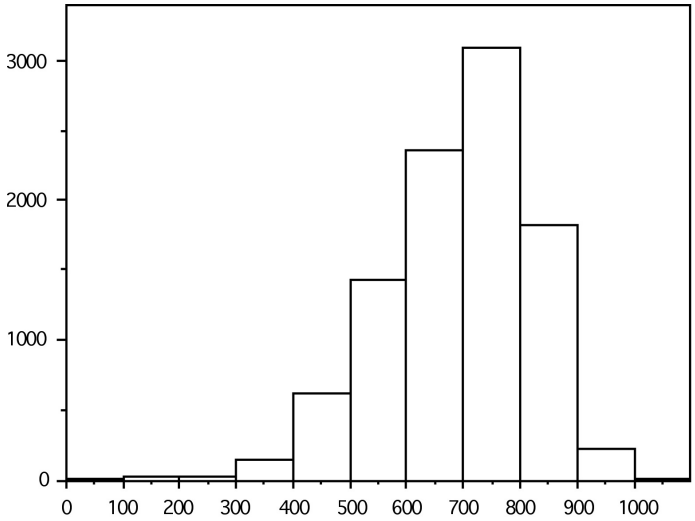


Abb. 23: Histogramm zur Verteilung der Längen der konservierten Sequenzbereiche der 8964 orthologen Promotorsequenzen. Mit dem Histogramm wird die Häufigkeitsverteilung der konservierten Promotersequenzen dargestellt. Hierzu werden die nach Größe geordneten Sequenzlängen in Klassen von 100 Basenpaare aufgeteilt. Die Häufigkeit mit der die Sequenzlängen jeder Klasse auftreten, wird mit Rechtecken, deren Fläche proportional zur klassenspezifischen Häufigkeit sind, dargestellt. Die Bereiche der konservierten Sequenzlängen kann auf der x-Achse abgelesen werden.

4.3.4 Untersuchung der konservierten humanen Sequenzbereiche auf potentielle TFBS

Die konservierten humanen Promotorsequenzbereiche werden auf potentielle TFBS mittels der im Internet verfügbaren Software MATCH™ gescannt. Dies erfolgt mit den in Tabelle 7 aufgelisteten Einstellungen:

Tab. 7: Einstellungen, die bei der Analyse mit dem Programm MATCH™ verwendet wurden

`group of matrices	`vertebrates
`high quality matrices´	YES
`core similarity´	0.85
`matrix similarity´	0.8

Die Eingabe der Sequenzen erfolgt im Fasta-Format. In den 8964 Promotorsequenzen werden 184 der 243 vorgegebenen TFBS gefunden. Insgesamt sind in den Promotorsequenzen 1.647.476 TFBS.

4.3.5 Korrektur der Anzahl der Bindungsstellen auf die Länge der konservierten Promotorsequenz

Durch die Reduzierung der Promotorsequenzen auf die konservierten Bereiche ist der Anteil an zufällig auftretenden funktionell nicht-relevanten TFBS verringert worden, eine vollständige Eliminierung erfolgt hierdurch jedoch nicht. Es ist zu erwarten, dass die Anzahl dieser zufällig auftretenden Bindungsstellen mit der Länge der konservierten Sequenz korreliert, d.h. je länger die konservierte Promotorsequenz ist, desto höher die Anzahl der Bindungsmotive. Bei einem rein funktionell bedingten Auftreten der TFBS ist eine solche Abhängigkeit nicht zu erwarten.
In Abbildung 24 ist die Abhängigkeit der Anzahl der gefundenen Bindungsmotive für die PWM V$MycMax_02, V$YY1_02 und V$Jun_01 in Abhängigkeit von der Länge der konservierten Promotorsequenzen dargestellt. Es wurden diese drei PWM hier exemplarisch ausgewählt, da V$MycMax_2 mit 19536 Bindungsstellen als Beispiel für eine PWM mit einer hohen Gesamtzahl von gefundenen Bindungsstellen, V$YY1_02 mit 7086 Bindungsstellen als Beispiel für eine PWM mit einer mittleren Gesamtzahl von Bindungsstellen und V$Jun_01 mit 866 Bindungsstellen als Beispiel für eine PWM mit einer geringen Gesamtzahl von Bindungsstellen steht. Bei V$MycMax_02 und V$YY1_02 ist eine deutliche Korrelation der Anzahl der potentiellen Bindungsstellen mit der Promotorlänge zu sehen. Bei V$Jun_01 ist die Anzahl der Bindungsstellen jedoch zu gering, um diese Tendenz deutlich erkennen zu können.

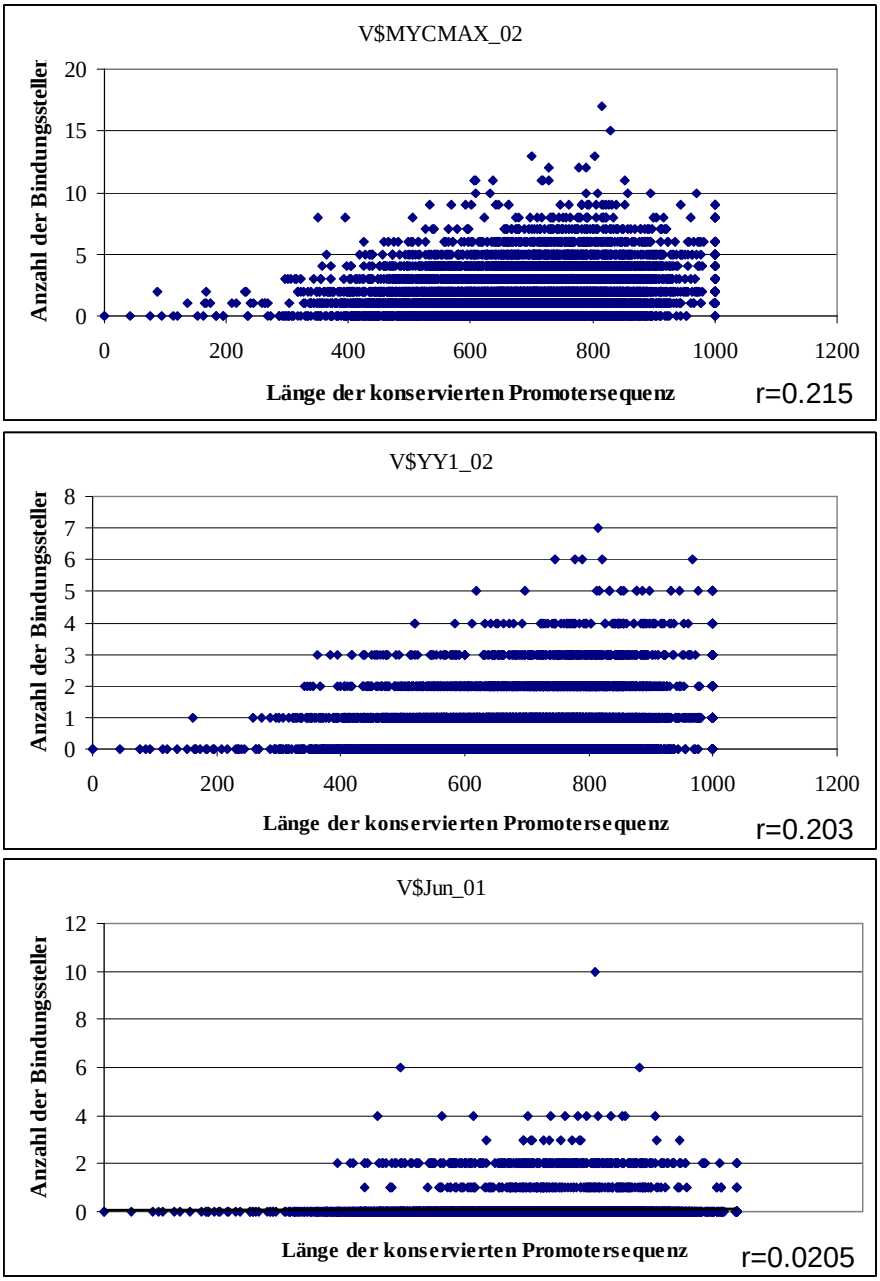


Abb. 24: Korrelation zwischen Länge der konservierten Promotersequenz und Anzahl der gefundenen Motive anhand von 3 Beispielen. In dem Scatterplot wird für die drei Beispiele V$MycMax_02, V$YY1_02 und V$Jun_01 der Zusammenhang zwischen der Länge der Promotorsequenzen auf der x-Achse und der Anzahl an gefundenen Motiven, auf der y-Achse, dargestellt. `r´ gibt den Korrelationskoeffizienten zwischen den beiden Parametern an. Der Korrelationskoeffizient ist ein Maß für den linearen Zusammenhang von zwei intervallskalierten Merkmalen. Er kann Werte zwischen -1 und 1 annehmen, Bei einem Wert von 1 besteht ein vollständig positiver bzw. bei einem Wert von -1 ein vollständig negativer linearer Zusammenhang. Bei einem Korrelationskoeffizienten von 0 besteht kein Zusammenhang zwischen den Variablen.Hiermit soll überprüft werden, ob die Anzahl der gefundenen Bindungsmotive von der Länge der konservierten Promotersequenz abhängig ist. Sowohl für V$MycMax_02, V$YY1_02 ist in der Graphik eine Korrelation zu sehen. Mit zunehmender Promoterlänge steigt die Anzahl der Bindungsstellen. Dies wird zudem durch die Korrelationskoeffizienten von r=0.215 und r=0.203 belegt. Bei V$Jun_01 ist die Korrelation nicht so deutlich,. Das ist vermutlich auf die geringe Gesamtzahl an Bindungsstellen zurückzuführen, die die Abschätzung einer Korrelation aufgrund einer geringen Zahl an Datenpunkten nicht zulässt.

Um zu überprüfen, ob diese Beobachtung auch für die anderen PWM gilt, wird in Abbildung 25 in einem Scatterplot jeweils die Gesamtzahl der gefundenen Bindungsstellen (über die 8964 Promotoren) gegen den Korrelationskoeffizient zwischen Promotorlänge und Anzahl pro Promotor gefunden Bindungsstellen dargestellt.

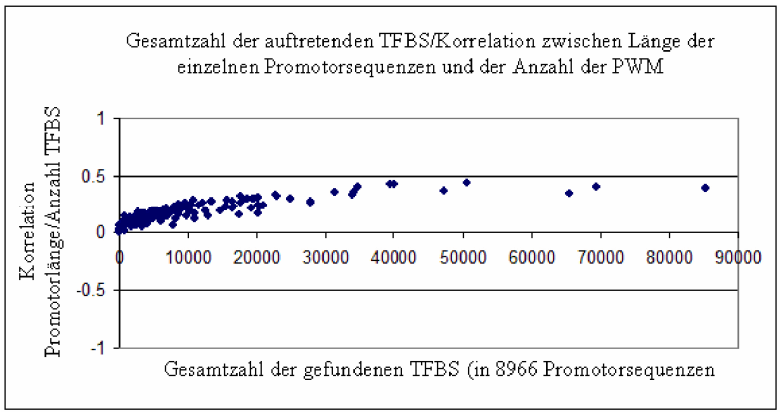


Abb. 25: Korrelation zwischen Länge der Promotersequenz und Anzahl der gefundenen Motive aller TFBS. Die Abbildung stellt das Verhältnis zwischen der Häufigkeit mit der eine TFBS in allen 8966 konservierten Promotorsequenzen auftritt und dem Korrelationskoeffizient zwischen Promotorlänge und Anzahl pro Promotor gefundenen TFBS dar. Mit zunehmender Gesamtzahl an gefundenen TFBS nimmt auch die Korrelation zwischen Promoterlänge und der Anzahl der TFBS zu. Somit gilt auch für andere TFBS, dass ihre Anzahl mit der Promoterlänge korreliert und dieser Effekt mit zunehmender Gesamtzahl des Auftretens deutlicher wird.

Abbildung 25 zeigt, dass diese Abhängigkeit auch für andere PWM gilt und dieser Effekt mit zunehmender Gesamtzahl deutlicher wird. Deshalb muss die Anzahl der für die einzelnen Promotoren gefundenen TFBS einer PWM für die jeweilige Sequenzlänge korrigiert werden. Die Korrektur erfolgt mittels einer z-Transformation. Mit der z-Transformation können beliebige (mindestens intervallskalierte) Verteilungen in eine neue Verteilung mit dem arithmetischen Mittel von Null und der Standardabweichung von 1, überführt werden. Durch diese Transformation wird ein Vergleich der Anzahl der Bindungsstellen auch bei unterschiedlicher Promotorlänge möglich, da durch die Umwandlung ein Sequenzlängen-unabhängiger standardisierter Wert erzeugt wird.

Die Promotorsequenzen werden anhand ihrer Längen in Untergruppen zusammengefasst, und für jede dieser Untergruppen und für jede PWM der jeweilige Mittelwert und die Standardabweichung der Anzahl der auftretenden Bindungsstellen berechnet.

Die z-Werte werden dann folgendermaßen berechnet:

$$z_i = \frac{x_i - \overline{x}}{s}$$

x_i = Anzahl der TFBS einer PWM in einer Promotorsequenz mit der Länge n

$\overline{x}$ = Mittelwert der Anzahl der TFBS einer PWM in allen Promotorsequenzen mit der Länge

s = Standardabweichung der Anzahl der TFBS einer PWM in allen Promotorsequenzen mit der Länge n

Es stellt sich die Frage, welcher Längenbereich für die Promotoren der Hintergrundgruppen gewählt werden müssen, um eine möglichst genaue Korrektur zu erzielen, d.h. um den Mittelwert und die Standardabweichung für jede PWM der jeweiligen Sequenzlängen möglichst genau abschätzen zu können. Dem gewählten Korrekturverfahren liegt die Annahme zugrunde, dass die Anzahl der vorhandenen TFBS unabhängig von der Länge der Promotorsequenz sein sollte und somit keine Korrelation zwischen der Promotorlänge und der Anzahl der TFBS (der verschiedenen Transkriptionsfaktoren) besteht. Damit stellt die „nicht mehr vorhandene“ Korrelation zwischen der Promotorlänge und der Anzahl der Bindungsstellen ein Maß für die erfolgreiche Korrektur dar.

Es wurden drei verschiedene Sequenzlängenbereiche für die Hintergrundgruppen getestet. So wurden die Promotorsequenzen in Gruppen eingeteilt, die Sequenzlängenbereiche von 20, 40 und 100 Nukleotiden umfassen. Die Hintergrundgruppe, die zum Beispiel zur Berechnung des z-Wertes einer Promotorsequenz mit 400 Nukleotiden fungiert, beinhaltet bei einer 20 Nukleotide umfassenden Hintergrundgruppe alle Promotorsequenzen des Gesamtsets, die 390 – 410 Nukleotide lang sind, bei der 40 Nukleotide umfassenden Hintergrundgruppe die Promotorsequenzen mit 380 – 420 Nukleotiden.

In Abbildung 26 wird die Verteilung der Korrelationskoeffizienten zwischen der Länge der Promotorsequenz und der Anzahl der TFBS vor der Korrektur, bzw. der z-Werte der verschiedenen PWM nach der Korrektur für die drei Hintergrundbereiche als Densityplots dargestellt. Alle drei Korrekturen reduzieren die Korrelation deutlich, der Ansatz, der Sequenzlängenbereiche von 20 Nukleotiden umfasst, zeigt nach der Transformation die geringsten Korrelationswerte.

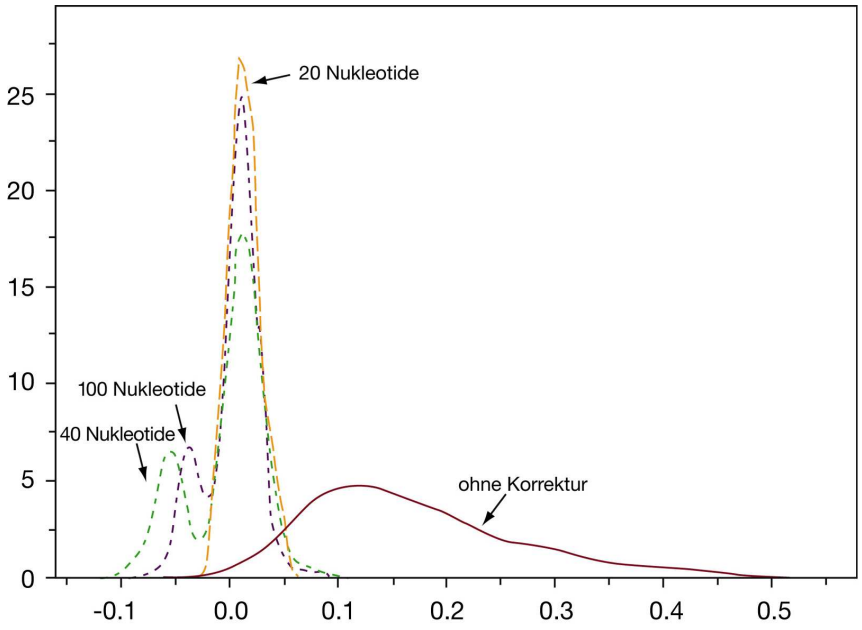


Abb. 26: Dichteplots der Verteilung der Korrelationskoeffizienten zwischen Promotorlängen und z-Wert ohne Korrektur und nach Korrektur anhand verschiedener Hintergrundsets. Mit dem Dichteplot wird die Verteilung der Korrelationskoeffizienten für die verschiedenen Gruppen dargestellt. Die Korrektur der Anzahl der jeweiligen Bindungsmotive erfolgt mit Hintergrundsets, die Sequenzen beinhalten die Sequenzlängen von 20, 40 bzw. 100 Nukleotiden. Durch eine Korrektur mit einem Sequenzlängenbereich von 20 wird die Korrelation zwischen Promotorlänge und Anzahl der gefundenen Motive am Effektivsten korrigiert, die Werte streuen am dichtesten um den Nullwert.

Die Abbildung 27 zeigt, wie sich die verschiedenen Korrekturen auf die Abhängigkeit der Korrelationskoeffizienten (Länge Promotorsequenzen/Anzahl TFBS) von der Gesamtzahl der einzelnen TFBS auswirken.

Bei allen drei Hintergrundbereichen wird die Korrelation deutlich reduziert.

Das Hintergrundset, das 20 Nukleotide umfasst, zeigt noch einen geringen Anstieg der Korrelationskoeffizienten mit zunehmender Gesamtzahl der Bindungsstellen. Insgesamt zeigt es aber die geringsten Korrelationskoeffizienten auf. Bei den Hintergrundsets, die 40 bzw. 100 bp umfassen, ist die Korrelation mit der Gesamtzahl weitgehend aufgehoben, aber die Werte streuen stärker um die Nullachse. Somit wurde die Korrektur anhand des 20 bp umfassenden Hintergrundsets durchgeführt.

Als Ergebnis der Korrektur ist eine Datenmatrix entstanden, die für jede Promotorsequenz und für jede PWM einen entsprechenden z-Wert enthält. Anhand dieser Matrix lassen sich nun die Werte aller Promotorsequenzen unabhängig von ihrer Länge untereinander vergleichen.

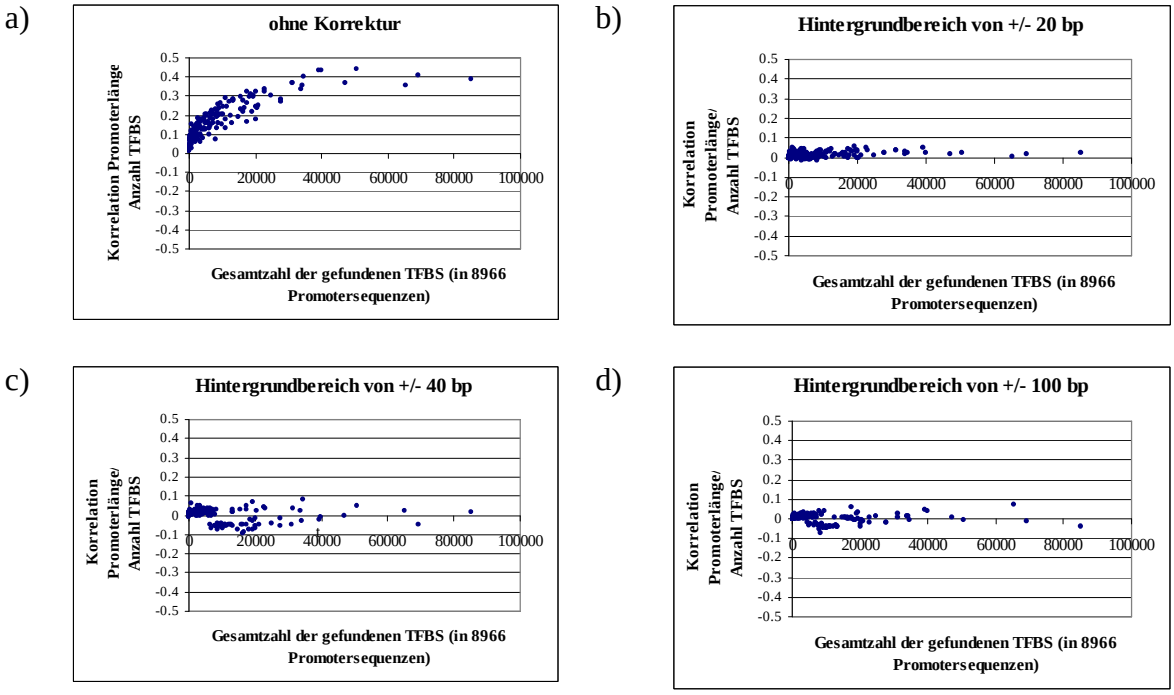


Abb. 27: Scatterplot der der Gesamtzahl der einzelnen TFBS und der Korrelationskoeffizienten zwischen Anzahl TFBS pro Promotor und der Promotorlänge nach der Korrektur mit verschiedenen Hintergrundsets. (a) Des nicht korrigierten Datensatzes (vergleiche Abbildung 25), (b) Korrektur anhand der 20 bp umfassenden Hintergrundsets, (c) Korrektur anhand der 40 bp umfassenden Hintergrundsets, (d) Korrektur anhand der 100 bp umfassenden Hintergrundsets.

4.3.6 Algorithmus zum Testen einer Gruppe ko-regulierter Gene auf die Anreicherung von TFBS

Das eigentliche Ziel der Methode ist die Untersuchung der Anreicherung von cis-regulatorischen Motiven in einer Gruppe von ko-regulierten Genen. Um dies zu testen, wird für jede PWM die Verteilung der z-Werte in der Gruppe ko-regulierter Gene mit der Verteilung in einem Hintergrundset, das aus den Promotorsequenzen aller auf dem Mikroarray vorhandenen Gene (für die eine Promotorsequenz vorliegt) besteht, verglichen. Dies erfolgt mit dem t-Test für zwei unabhängige Stichproben.

Die Nullhypothese wird folgendermaßen formuliert:
Wenn innerhalb der Hintergrundliste eine hohe Anzahl von Promotoren sind, die eine bestimmte TFBS haben, werden sich innerhalb der Promotorsequenzen der ko-regulierten Gene, auch wenn diese nur zufällig aus der Gesamtliste ausgewählt wurden, eine entsprechend hohe Anzahl von Promotorsequenzen mit der TFBS finden.

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Der t-Wert wird wie folgt berechnet:

$$t_{n_1+n_2-2} = \frac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}_{\overline{x}_1-\overline{x}_2}}$$

$\overline{x}_1 - \overline{x}_2$ = Differenz der Mittelwerte der beiden Gruppen

$\hat{\sigma}_{\overline{x}_1-\overline{x}_2}$ = Differenz der Standardabweichungen der beiden Gruppen

Die Signifikanz der t-Statistik wurde mittels eines Permutationsverfahrens bewertet. Hierbei wird für jede PWM die berechnete t-Statistik mit den t-Statistiken von 10000 zufällig ausgewählten Gensets gleicher Größe aus dem Hintergrundset verglichen. Die Bestimmung des p-Wertes erfolgt, indem die Anzahl der Gensets, welche eine höhere t-Wert haben als die getestete Gruppe ko-regulierter Gene, durch die entsprechende Zahl von Permutationen, hier 10000, dividiert wird.
Es wird das Permutationenverfahren verwendet, da dies für Datensätze, die keiner Normalverteilung folgen, bzw. für kleine Datensätzen genauer ist, als die Bestimmung des p-Wertes über die Verteilung der t-Werte.

Da eine Vielzahl von PWM parallel getestet werden, liegt ein multiples Testproblem vor, für welches korrigiert werden muss (siehe 2.2.9). Hierzu wurde das Verfahren der Step-up Prozedur zur Kontrolle der *false discovery rate* (FDR) nach Verfahren von Benjamini & Hochberg verwendet (Hochberg and Benjamini 1990). Dies kontrolliert die FDR auf dem Level α, wenn die p-Werte aus der Null-Verteilung unabhängig und gleich verteilt sind.

4.3.7 Analyse eines Datensatzes

Mit der Auswertung des folgenden Datensatzes soll gezeigt werden, dass die Methode für die Identifizierung von cis-regulatorischen Elementen geeignet ist.

Datensatz

Es handelt sich hierbei um den in Kapitel 3.4.1 beschriebenen Datensatz, der bereits für die Darstellung verschiedener anderer Metaanalysemethoden verwendet wurde.

In Tabelle 8 sind die Ergebnisse der Auswertung dargestellt. Insgesamt haben 18 PWM eine *false discovery rate* kleiner 25% (Benjamini and Hochberg 1995). Bei den am stärksten angereicherten Motiven handelt es sich um PWM die E-Box Motive d.h. die Sequenz CACGTG enthalten. Unter den 18 PWM sind 10 E-Box bzw. E-Box-ähnliche Motive. Neben den E-Box Motiven findet sich auf Rang 12 die PWM für das Bindungsmotiv des Transkriptionsfaktor YY1. Die Interaktion von c-Myc und YY1 wurde 1993 bzw. 1996 von Riggs, Saleque et al. (1993) und Shrivastava, Yu et al. (1996) publiziert. Sie zeigten, dass überexprimiertes YY1 die Expression von sowohl endogenem als auch exogenem c-myc aktivieren kann. Zudem verändert die Überexpression von c-myc die konstitutive Repression durch YY1, indem es die Bindung zwischen YY1 und den basalen Transkriptionsproteinen wie dem TATA-Bindungsprotein und dem Transkriptionsfaktor IIF verhindert (Austen, Cerni et al. 1998). Hieraus würde man die Anreicherung des Bindungsmotivs für YY1 erwarten.

Tab. 8: Ergebnis der Anreichung von cis-regulatorischen Motiven einer Liste von ko-regulierten Genen, die bei dem Vergleich der Genexpression von T-Zellen von Eμ MYC-wt Mäusen und Eμ-Mäusen herauf reguliert wurden.

PWM (TRANSFAC)	t-Statistik	p-Wert	FDR	Sequenz der PWM
V$ARNT_01	3.16	0.0001	0.0175	NDDNNCACGTGNNNNN
V$MYCMAX_01	3.023	0.0004	0.028	NNACCACGTGGTNN
V$MAX_01	3.107	0.0005	0.028	NNANCACGTGNTNN
V$USF_01	2.936	0.0008	0.028	NNRYCACGTGRYNN
V$NMYC_01	2.85	0.0008	0.028	NNCCACGTGNNN
V$ER_Q6	2.742	0.0011	0.0320833	NNARGNCANNNTGACCYNN
V$MYCMAX_02	2.738	0.0026	0.065	NANCACGTGNNW
V$T3R_01	2.531	0.004	0.0875	SNNTRAGGTCACGSNN
V$AREB6_03	2.607	0.0053	0.09275	VNRCACCTGKNC
V$R_01	2.494	0.0071	0.09275	NNGKCCNCSNRNYGTGGTGCN
V$RORA1_01	2.253	0.0086	0.1129546	NWAWNNAGGTCAN
V$YY1_02	2.237	0.0101	0.1254167	NNNCGGCCATCTTGNCTSNW
V$CREB_01	2.453	0.0109	0.1295	TGACGTMA
V$EGR3_01	1.986	0.0111	0.1295	NTGCGTGGGCGK
V$CREBP1_Q2	2.219	0.0238	0.1295	VGTGACGTMACN
V$USF_C	2.251	0.024	0.1815625	NCACGTGN
V$XBP1_01	1.986	0.0292	0.2333333	NNGNTGACGTGKNNNWT
V$CREL_01	2.086	0.0295	0.2333333	SGGRNTTTCC

Tab. 9: Ergebnis der Anreichung von cis-regulatorischen Motiven einer Liste von ko-regulierten Genen, die bei dem Vergleich der Genexpression in T-Zellen von Eμ MYC-wt Mäusen und Eμ-Mäusen herunterreguliert wurden

PWM	t-Statistik	p-Value	FDR	Sequenz der PWM
V$HFH3_01	1.848	0.0171	0.588	KNNTRTTTRTTTA
V$HFH8_01	1.819	0.0196	0.588	NNNTGTTTATNTR
V$HOX13_01	1.624	0.0222	0.588	TGCNHNCWYCCYCATTAKTNNDCNMNHYCN
V$HNF3B_01	1.616	0.0229	0.588	KGNANTRTTTRYTTW
V$POLY_C	1.524	0.0259	0.588	CAATAAAACCYYYYKCTN
V$BRN2_01	1.613	0.0272	0.588	NNCATNSRWAATNMRN
V$GATA_C	1.6	0.0298	0.588	NGATAAGNMNN
V$GATA1_03	1.685	0.0343	0.588	ANGNDGATAANNGN
V$ZID_01	1.817	0.0347	0.588	NGGCTCYATCAYC
V$CHOP_01	1.881	0.0411	0.588	NNRTGCAATMCCC
V$PAX5_01	1.417	0.0497	0.588	BCNNNRNGCANBGNTGNRTAGCSGCHNB
V$PAX5_02	1.133	0.0621	0.588	RRMSWGANWYCTNRAGCGKRACSRYNSM
V$COUP_01	1.444	0.0683	0.588	TGAMCTTTGMMCYT
V$HAND1E47_01	1.514	0.0729	0.588	NNNNGNRTCTGGMWTT
V$LMO2COM_02	1.328	0.0839	0.588	NMGATANSG
V$GATA1_02	1.322	0.0916	0.639	NNNNNGATANKGNN

In Tabelle 9 ist das Ergebnis der Promotoranalyse der Gene, die durch c-Myc herunterreguliert werden. Hier ist keine der PWM signifikant angereichert.

4.3.8 Diskussion

Im Rahmen dieses Kapitels ist die Etablierung einer Methode zur Identifizierung von funktionell relevanten cis-regulatorischen Motiven in einem Set von ko-regulierten Genen beschrieben worden.

Hierzu wurden

- orthologe Human – Maus –Promotoren aus der *CSHLmpd* Datenbank heruntergeladen
- repetitive Sequenzbereiche mittels des Programms *Repeatmasker* maskiert
- konservierte Sequenzbereiche durch ein Alignment des Programmes *DiAlign-2.*2 extrahiert
- potentielle Bindungsmotive mit $MATCH_{TM}$ identifiziert

- die Anzahl der Bindungsstellen auf die Länge der konservierten Promotorsequenz durch eine z-Transformation korrigiert
- die Anreicherung von TFBS in ko-regulierten Genen berechnet

Die Methode wurde an einem Mikroarray-Datensatz getestet, in dem die Genexpression der T-Zellen von Eμ MYC-WT Mäusen und Eμ-Mäusen verglichen wurde. Es hat sich gezeigt, dass die Methode zur Identifizierung von cis-regulatorischen Elementen geeignet ist. So wurde in der Liste der Gene, die in Eμ MYC-wt Mäusen höher exprimiert waren, eine sehr deutliche Anreicherung von E-Boxen und E-Box-ähnlichen Motiven gefunden. Auch findet sich YY1, welcher als Ko-Aktivator von c-Myb beschrieben wurde. In der Liste der herunterregulierten Gene finden sich keine signifikant angereicherten TFBS. Hier spiegelt sich möglicherweise wider, dass es für die Repression durch c-Myc verschiedene Mechanismen gibt und damit keiner der Mechanismen deutlich hervortritt.

Als Basis für das Alignment orthologer Promotorsequenzen wurden humane und Mauspromotoren gewählt. Die Verwendung von Promotorsequenzen von mehr als zwei Organismen hätte zwar zu einer Erhöhung des Konservierungsgrades und vermutlich zu einer besseren Trennung von funktionell relevanten zu funktionell nicht relevanten Sequenzmotiven geführt, gleichzeitig aber zur Reduzierung der Anzahl der zur Verfügung stehenden orthologen Sequenzen und damit zu einer Reduzierung der statistischen Aussagekraft bei der Berechung von signifikant angereicherten TFBS.
Die humanen Promotorsequenzen und die dazu gehörigen orthologen Mauspromotoren wurden aus der CSHLmpd Datenbank heruntergeladen. Diese Datenbank wurde gewählt, da sie eine sehr große Zahl sehr gut annotierter Promotoren beinhaltet. So beinhaltet sie u.a. die Promotorsequenzen der Datenbanken DBTSS und der EPD. Beide Datenbanken enthalten nur die experimentell bestimmten Promotorsequenzen. Die der DBTSS zugrunde liegende *oligo-capping* Methode verbessert zudem die genaue Lokalisation der Promotoren. Für den Erfolg des phylogenetischen *Footprintings* ist entscheidend, dass die orthologe Beziehung zwischen den zu vergleichenden Sequenzen vorhanden ist.

Der hohe Konservierungsgrad, der im Mittel 69,5% beträgt, entspricht annähernd dem von Levy et al. (2002) gefundenen Wert, die einen mittleren Konservierungsgrad von 75% zwischen orthologen Mensch- und Mauspromotorsequenzen angeben.
Als Sequenzbereich wurde -700 bis +300 bp um die TSS gewählt. Dieser Bereich wurde gewählt, da hierfür der höchste Konservierungsgrad im Vergleich zu längeren bzw. kürzeren Sequenzbereichen festgestellt wurde, was vermuten lässt, dass hier die höchste Anreicherung von konservierte TFBS vorliegt (Suzuki, Yamashita et al. 2004).

Für jedes Gen wurde ein Promotor, der als *„best“* klassifiziert wurde, aus der Datenbank heruntergeladen. Die Verwendung von nur einem Promotor für jedes Gen stellt jedoch eine Einschränkung für die Methode dar. Nahezu die Hälfte der Kernpromotoren haben nicht nur eine einzige TSS, sondern eine Anordnung von räumlich benachbarten TSS, von denen die Transkription mit unterschiedlicher Rate initiiert wird (Frith, Li et al. 2003). Dies ermöglicht eine weitere Diversifizierung der Transkription innerhalb eines Promotors, indem durch die Verwendung verschiedener TSS unterschiedliche Module transkriptionell regulatorischer Elemente verwendet werden können (Landry et al. 2003; Carninci et al. 2005; Cheng et al. 2005; Kim et al. 2005; Kimura et al. 2006). Es wird angenommen, dass in dieser Diversität ein Schlüssel für die Entwicklung von hoch komplexen Systemen wie z.B. dem Gehirn oder dem Immunsystem liegt und begründet, warum die Gesamtzahl der humanen Gene, die auf 20,000–25,000 geschätzt wird (International Human Genome Sequencing Consortium, 2004), sich nicht so sehr von der Zahl der Gene von Hefe, Fliegen und Würmern unterscheidet (Goffeau et al. 1996; C. elegans Sequencing Consortium 1998). Dies gilt auch im Vergleich zu anderen Mammalia, wie Maus, Hund und Kuh, da der Mensch deutlich weiter entwickelte physiologische, anatomische und metabolische Charakteristika hat, obwohl auch hier die Zahl der Gene vergleichbar ist. Es sind eine Reihe von Publikationen erschienen, die Spezies-spezifische Charakteristika in Hinsicht auf Signaltransduktion, Reaktionen auf Wachstumsfaktoren, neuronale Verbindungen und Medikamentenmetabolismus auf alternative Promotoren und auch alternative Splicevarianten zurückführen (Grandien et al. 1997; Luzi et al. 2000; Tautz 2000; Dermitzakis and Clark 2002; Su and Gladyshev 2004; Pan et al. 2005; Wu 2005). Diese Annahme stellt für die Verwendung des phylogenetischen *Footprinting* eine deutliche Limitierung dar. Wenn alternative

Promotoren Spezies-spezifisch auftreten, könnten sie anhand dieser Methode nicht eindeutig detektiert werden.
Die Existenz von alternativen Promotorbereichen ist zudem eine zusätzliche Herausforderung für die Identifizierung und Lokalisation der Promotoren.

Inzwischen wurde in einer Reihe von Publikationen beschrieben, dass viele TF in Bereichen binden, die weit entfernt von dem annotierten Gen liegen. So konnte z.B. mit *Tiling Arrays*, die das Chromosom 21 und 22 abdecken, gezeigt werden, dass nur ein geringer Teil der p65 bzw. NF-KappaB Bindungsstellen auf dem Chromosom 22 in der Nähe des Promotors auftreten (Martone, Euskirchen et al. 2003). Ähnliches wurde für den Östrogenrezeptor auf Chromosome 21 und 22, und für CREB und STAT1/2 auf dem Chromosome 22 beschrieben (Martone, Euskirchen et al. 2003; Euskirchen, Royce et al. 2004; Hartman, Bertone et al. 2005; Carroll, Meyer et al. 2006). Über die Funktionsweise und Bedeutung dieser Bindungsstellen ist bisher nur wenig bekannt. Für distal liegende Östrogen-Rezeptorbindungsstellen wurde gezeigt, dass sie z.B. als Enhancer fungieren können (Carroll, Meyer et al. 2006). Zudem wurde ihre Bedeutung bei der Regulation von nicht-kodierender RNA gezeigt (Cawley, Bekiranov et al. 2004). Die Erfassung von konservierten Bindungsmotiven außerhalb des bisher verwendeten Promotorbereiches sollte bei einer möglichen Erweiterung und Optimierung der Methode berücksichtigt werden.
Eine Anforderung an das hier verwendete Alignment-Programm war, dass es auch bei Organismen, die einen geringen phylogenetischen Abstand haben wie dem Mensch und der Maus, zwischen konservierten und nicht-konservieren Sequenzbereichen unterscheiden kann. Mittels des Alignments sollten vor allem kurze Sequenzmotive, deren Reihenfolge innerhalb der Sequenz sich auch verändert haben kann, gefunden werden. Das Alignmentprogramm *DiAlign-2.2* wurde gewählt, da es sowohl eine lokales als auch ein globales Alignment durchführt. Somit kann es lokale Ähnlichkeiten, die räumlich durch unverwandte Sequenzen getrennt sind, aber auch Sequenzbereiche, in denen funktionelle Elemente ihre Positionen getauscht haben, erkennen.

Laut Sauer, Shelest et al. (2006) ist die Auswahl des Alignment-Algorithmus nicht von so großer Bedeutung für das Ergebnis. Sie haben verschiedene Alignment-Methoden verglichen und keine signifikanten Unterschiede in den Ergebnissen der verschiedenen

Programme gefunden. Dies gilt vor allem für das Alignment von Spezies mit einer nur geringen evolutionären Distanz. Auch Pollard, Bergman et al. (2004) haben verschiedene Alignment-Programme getestet. Sie zeigten dagegen, dass *DiAlign-2* beim Vergleich von nicht-kodierenden Human-/Maussequenzen bessere Ergebnisse erzielte als andere Alignment-Programme.

Die hier gezeigte Methode beinhaltet u.a. das Auffinden von Motiven, die durch PWM beschrieben werden. Dies erfolgte mit dem web-basierten Programm $MATCH_{TM}$, das nach TFBS anhand von Motivmustern der TRANSFAC® Datenbank sucht. Die Selektion von Motiven innerhalb der Sequenzen als potentielle TFBS erfolgt anhand des *matrix similarity scores* und des *core similarity scores*. Als Schwellenwert sind die Werte 0.85 für den *core similarity score* und 0.8 für den *matrix similarity score* gewählt worden. Es wurden bewusst Schwellenwerte gewählt, die häufig niedriger als die vom Programm vorgegebenen optimalen Schwellenwerte liegen. Dies beruht auf der Annahme, dass die Erstellung der PWM anhand experimentell nachgewiesener TFBS erfolgt und somit die Genauigkeit der PWM von der Anzahl der beschriebenen TFBS abhängt, aber auch von der Art der gewählten Experimente, welche zur Identifizierung dieser TFBS geführt haben. Basieren die Motive z.B. vorrangig auf Experimenten, in denen eine aktivierende Funktion eines Transkriptionsfaktors eine Rolle spielt, wird das gefundene Motiv einer Bindungsstelle entsprechen, die bei einer Aktivierung verwendet wird. Es ist jedoch möglich, dass der Transkriptionsfaktor in einem anderen funktionellen Kontext an ein leicht variierendes Motiv bindet. Durch die Wahl von geringeren Schwellenwerten sollte eine höhere Flexibilität für das Auffinden solcher möglichen Bindungsmotive geschaffen werden.

Ein wichtiges Element ist die Korrektur der Anzahl der Bindungsstellen in Bezug auf die Länge der konservierten Promotoren. Durch die z-Transformation wurden Sequenzlängen-unabhängige, standardisierte Werte erzeugt und so der Zusammenhang zwischen Promotorlänge und Anzahl der Bindungsstellen aufgehoben.

Die Promotorsequenzen wurden auf 243 PWM mit $MATCH_{TM}$ untersucht. Von diesen wurden nur 169 in den 8964 Promotorsequenzen gefunden.

Die Verwendung von PWM, wie sie in TRANSFAC® vorliegen, führt zu einer Reduzierung der falsch negativen Bindungsmotive gegenüber der Konsensussequenz.

Xue, Wang et al. (2004) haben das Auftreten von Bindungsstellen anhand von TRANSFAC® Matrizen in stromaufwärts der TSS liegenden Sequenzen mit dem Auftreten in Exon 2 verglichen. Sie fanden, dass weniger als 50% der TRANSFAC® Matrizen in den stromaufwärts liegenden Sequenzen angereichert werden. Diese niedrige Auffindungssrate erklären sie mit einer geringen Spezifität der Matrizen. Die geringe Spezifität kann darauf beruhen, dass die TFBS zu kurz sind oder dass sie auf einer nur geringen Anzahl von experimentell bestimmten Bindungsstellen beruhen, die die Erstellung eines genauen Modells nicht ermöglichen. TFBS, die eine geringe Spezifität aufweisen, sind mit der vorliegenden Methode nur schwierig zu finden – auch wenn die Schwellenwerte niedrig gewählt wurden.

Neben einer geringen Spezifität gibt es eine Reihe von Gründen warum TFBS mittels des phylogenetischen *Footprintings* nicht gefunden werden.

- TFBS sind Art-spezifisch und werden nicht als orthologe Sequenz identifiziert.
- TFBS können z.B. Insertionen oder Deletionen aufweisen, die im Falle einer noch funktionierenden Bindungsstelle nicht auffallen.
- Einige TF binden als Dimere. Diese TFBS erscheinen als zwei konservierte Regionen, die nur von wenigen variablen Nukleotiden getrennt werden. Aufgrund der Variabeln zwischen den Sequenzen kann diese Bindungsstelle nur schwierig detektiert werden.
- Einige TF (z.B. Sp1 and C/EBP) binden an sehr variierende Sequenzen. Der evolutionäre Druck, diese Bindungsstellen zu erhalten, ist sehr gering, da die Wahrscheinlichkeit einer ähnlichen Bindungsstelle in der Nachbarschaft hoch ist.
- Neben dem Sequenzmotiv werden die Bindungseigenschaften von TFBS von anderen Faktoren wie der DNA Struktur und das Binden von anderen TF beeinflusst.
- Der Wert der PWM liegt unterhalb des Schwellenwertes. Dies kann z.B. erfolgen, wenn mehrere TF als Modul kooperativ binden und einzelne TF nur eine schwache Bindung haben, aber in Kombination mit anderen TF eine stabile Bindung eingehen. Einige Forscher haben die Eigenschaft von TF, gemeinsam zu binden, in die computerbasierte Vorhersage von transkriptionellen Netzwerken bereits integriert (Zhu, Pilpel et al. 2002).

Die Anreicherung von cis-regulativen Elementen wurde durch einen ungepaarten permutationsbasierten T-Test berechnet. Dieses Verfahren ermöglicht, auch bei nicht normal verteilten Daten bzw. kleinen Datensätzen ein Signifikanzlevel zu berechnen. Eine Situation, die vor allem für TFBS, die nur selten auftreten, wichtig ist. Zu berücksichtigen ist, dass bei diesem Ansatz die mögliche Korrelation einzelner Gene nicht berücksichtigt wird, d.h. auch wenn keinerlei Effekt für einen TFBS vorliegt, kann dennoch eine andere Verteilung gefunden werden als im Hintergrundset, da Gene aufgrund eines anderen als des beobachteten Phänotyps korrelieren können. Hieraus können in der Ergebnisliste falsch positive TFBS resultieren.

Mittels der vorliegenden bioinformatischen Methode konnten erfolgreich angereicherte TFBS in einen Set von ko-regulierten Genen identifiziert werden. Es wurden einige Einschränkungen aufgezeigt, z.B. alternative Promotoren, fehlende orthologe Beziehungen zwischen den Organismen, Bindungsstellen, die außerhalb des Promotorbereiches liegen, die bei der Interpretation der Ergebnisse berücksichtigt werden müssen. Dennoch ist sie ein grundlegender Schritt für die Identifizierung von regulatorischen Zusammenhängen, die dann in Kontext mit anderen regulatorischen Mechanismen, z.B. der miRNA abhängigen Regulation, gesetzt werden, bzw. als Basis für einen konkreten experimentellen Ansatz zur Untersuchung der Genregulation innerhalb eines bestimmten Phänotypes fungieren kann. Die Bedeutung der Kombination von Computer basierten und experimentellen Ansätzen wurde immer wieder aufgezeigt. So konnte z.B. mittels des experimentellen Ansatzes der *oligo-capping* Methode (Suzuki et al, 2002) eine bessere Lokalisierung der Promotorbereiche erzielt werden. Auch zeigt sich die Notwendigkeit, PWM anhand weiterer, experimentell bestimmter Bindungsstellen genauer zu beschreiben. Ein zusätzlicher Fokus sollte hierbei auf die unterschiedlichen Kontexte, in denen die Transkriptionsfaktoren agieren, gelegt werden, um so eine mögliche kontextabhängige Variabilität zu erfassen.
Die Literatur der letzten Jahre zeigt, dass trotz aller Fortschritte die Detektion von DNA Bindungsmotiven in höheren Organismen immer noch eine Herausforderung für Biologen und Bioinformatiker bleibt.

5 Etablierung der statistischen Auswertung einer miRNA-Mikroarray Plattform

5.1 Einleitung

50 Jahre lang galt als das „zentrale Dogma der Molekularbiologie“, dass **ein** Gen über die Bildung **einer** mRNA in **ein** Protein übersetzt wird.

Abb. 28: „Zentrales Dogma der Molekularbiologie“

Diese Annahme kann heute nur noch für Bakterien aufrecht erhalten werden. Das Genom von höheren Organismen ist weit vielfältiger als ursprünglich angenommen. Neben der protein-kodierenden RNA findet sich hier auch nicht-kodierende RNA, deren Anteil mit zunehmender Komplexität der Organismen steigt. Es wird ein Zusammenhang zwischen dem Anteil der nicht-kodierenden RNA und der Komplexität der Organismen angenommen (Mattick 2003). Hierfür spricht auch, dass sich die Anzahl der Gene mit zunehmender Komplexität nicht deutlich erhöht. So hat der Mensch mit ca. 30000 Genen nur etwa doppelt so viele Gene wie *Drosophila melanogaster* und nur ca. 10000 Gene mehr als der Wurm *Caenorhabditis elegans* (Mattick 2003). Während in der Vergangenheit diese nicht-kodierenden RNA-Transkripte als *cloning artifacts* oder *truncated molecules* bezeichnet wurden, weiß man heute, dass sie wichtige biologische Funktionen übernehmen. Die wichtige funktionelle Bedeutung dieser Transkripte zeigt sich auch darin, dass ihre Promotorbereiche im allgemeinen besser evolutionär konserviert sind als die Promotoren der protein-kodierenden RNAs (Ravasi, Suzuki et al. 2006).

Nicht-kodierende RNAs (ncRNAs) können in verschiedene Gruppen unterschieden werden. Hierzu gehören:

- ribosomale RNAs (rRNA)
- transfer-RNAs (tRNA)

- small nucleolar RNAs (snoRNA)
- mikroRNAs (miRNA)

Die verschiedenen Gruppen übernehmen unterschiedliche Aufgaben in der Zelle. Während die rRNA zusammen mit den ribosomalen Proteinen am Aufbau und der enzymatischen Aktivität des Ribosoms beteiligt ist, vermittelt die tRNA bei der Translation die richtige Aminosäure zum entsprechenden Codon auf der mRNA. SnoRNA sind an der Prozessierung und Modifikation anderer Ribonukleinsäuren – insbesondere ribosomaler RNA – beteiligt (Brown 2002). MiRNAs haben genregulatorische Funktionen.

In Abbildung 29 wird der Einfluss, den verschiedene *non-coding* RNAs z.B. auf die Genexpression haben können, dargestellt.

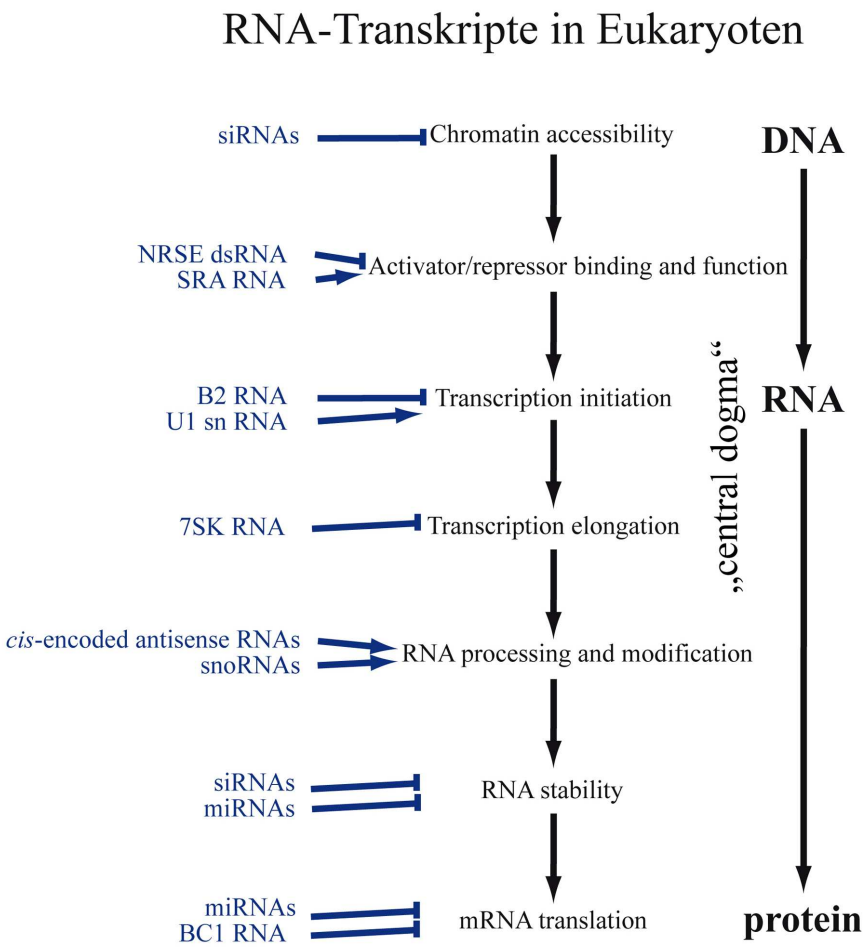


Abb. 29: Schritte in der Genregulation, die durch ncRNAs beeinflusst werden (modifiziert nach Storz, Altuvia et al. (2005)).

Auch in Prokaryoten wurden ncRNAs gefunden (Vogel, Bartels et al. 2003; Gottesman 2004; Storz, Opdyke et al. 2004; Wilderman, Sowa et al. 2004; Dennis and Omer 2005; Kawano, Reynolds et al. 2005; Mattick and Makunin 2006). Sie spielen hier aber nur eine untergeordnete Rolle (Mattick 2004).

5.1.1 miRNAs

Die mikroRNAs wurden erstmals 1993 in *Caenorhabditis elegans* als Regulatoren von Entwicklungsstadien beschrieben (Lee, Feinbaum et al. 1993). Seit ihrer Entdeckung wurde ihre Bedeutung als Regulatoren für eine Vielzahl von Prozessen erkannt. Hierzu gehören z.B. zeit- und gewebespezifische miRNA-Expressionsmuster bei der Entwicklung von Pflanzen und Tieren (Lau, Lim et al. 2001; Lagos-Quintana, Rauhut et al. 2002; Aravin, Lagos-Quintana et al. 2003; Houbaviy, Murray et al. 2003; Lim, Lau et al. 2003; Ambros 2004; Bartel 2004; He and Hannon 2004) oder die Regulation von physiologischen Prozessen wie z.B. Apoptose, Zellteilung und Zelldifferenzierung (Volinia, Calin et al. 2006). Auch wurde die Existenz von tumortyp-spezifischen miRNA Expressionsprofilen z.B. für die chronisch lymphatische Leukämie, das kolorektale Adenokarzinom, das Burkitt-Lymphom, das Glioblastom und das Lungenkarzinom gezeigt. Zum Teil konnten die Expressionsmuster auch zur Klassifizierung von Tumorsubtypen verwendet werden (Michael, SM et al. 2003; He and Hannon 2004; Metzler, Wilda et al. 2004; Calin, Ferracin et al. 2005; Chan, Krichevsky et al. 2005; Ciafre, Galardi et al. 2005; Croce and Calin 2005; Gregory and Shiekhattar 2005; Johnson, Grosshans et al. 2005; Lu, Qian et al. 2005; Cummins and Velculescu 2006; Volinia, Calin et al. 2006).

miRNAs zeigen einen hohen phylogenetischen Konservierungsgrad. So sind konservierte Sequenzen sowohl bei *Caenorhabditis elegans*, *Drosophila melanogaster* als auch beim Menschen zu finden. (Pasquinelli, Reinhart et al. 2000; Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001; Lai, Tomancak et al. 2003; Lim, Lau et al. 2003).

Die aktive reife miRNA besteht aus 17 - 24 Nukleotide langen einzelsträngigen RNA-Molekülen. Sie wird in zwei Schritten prozessiert.

Wie in Abbildung 30 dargestellt, wird zunächst die sogenannte pri-miRNA im Zellkern durch die RNA-Polymerase II transkribiert (Lee, Kim et al. 2004). Die Sequenzbereiche

für die pri-miRNA Trankripte liegen sowohl in den Intergen-Regionen als auch im Intron-Bereich der pre-mRNAs, was zu einer gemeinsamen Transkription mit den entsprechenden Genen führt. Das Enzym Drosha, eine TypIII RNAse, prozessiert die pri-miRNA zur 50–70 Nukleotid langen pre-miRNA. Die ~70 Nukleotid lange pre-miRNA bildet eine Haarnadel–Struktur aus, welche dann mittels Exportin-5 aus dem Zellkern in das Zytoplasma transportiert wird.

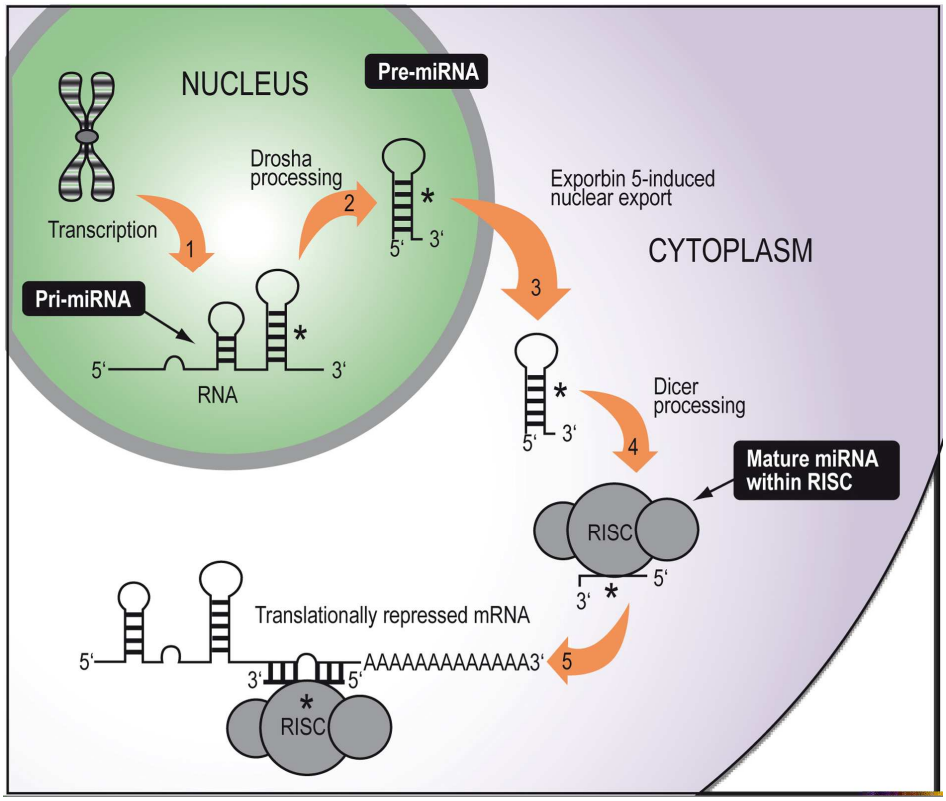


Abb. 30:Prozessierung der miRNA (http://www.ambion.com). Die pri-miRNA wird im Zellkern durch die RNA-Polymerase II transkribiert und anschließend mit dem Enzym Drosha in 50-70 Nukleotide lange pre-miRNA Fragmente prozessiert. Diese ~70 Nukleotid lange pre-miRNA bildet eine Haarnadel-Struktur aus, die durch Exportin-5 aus dem Zellkern in das Zytoplasma transportiert wird. Dort erfolgt die weitere Prozessierung zum reifen miRNA-Duplex (18 – 24 Nukleotide), indem das Enzym Dicer die Loop-Strukur der pre-miRNA entfernt.

Dort entfernt das Enzym Dicer die Loop-Struktur der pre-miRNA, die dadurch zur reifen miRNA-Duplex (18 – 24 Nukleotide) prozessiert wird. In einem Komplex mit RISC (*RNA-induced silencing complex*) wird die doppelsträngige miRNA in die einzelsträngige Form überführt. Die bisher bekannte Funktion von miRNAs liegt in der Antisense Regulation von mRNA-Transkripten. So binden miRNAs im 3´-UTR Bereich ihrer Ziel-mRNA, was bei einer vollständigen Komplementarität zur Hemmung der Translation und bei einer teilweisen Komplementarität zum Abbau der Ziel-mRNA

führt. Es wird angenommen, dass eine miRNA mehrere hundert Ziel-mRNAs haben kann, aber mehrere miRNAs können auch an die gleiche Ziel-mRNA binden. Das humane Genom soll ca. 1000 verschiedene miRNAs enthalten, welche die Expression von mindestens 30% der humanen Proteine regulieren (Lewis, Burge et al. 2005). miRNAs sind somit ein wichtiger Bestandteil von genregulatorischen Netzwerken und damit von elementarer Bedeutung bei der Untersuchung von biologischen und medizinischen Fragestellungen. Da häufig eine Vielzahl von miRNAs an der Regulation eines Prozesses beteiligt sind, ist es notwendig, die Expression der verschiedenen miRNAs gleichzeitig zu messen, um so entsprechende Expressionsmuster zu finden. Zur gleichzeitigen Untersuchung verschiedener miRNAs eignen sich miRNA-Mikroarrays, deren Technologie und Auswertung am IMT etabliert werden sollte. Da es sich hierbei um einen hochdimensionalen Datensatz handelt, müssen die bereits in Kapitel 2 beschriebenen Aspekte z.B. hinsichtlich

- experimentellem Design
- Bildanalyse
- Qualitätskontrollen
- Normalisierung
- Selektion differentiell exprimierter miRNAs

berücksichtigt werden. Die in Kapitel 2 beschriebenen statistischen Methoden werden somit für die Auswertung des Datensatzes verwendet. Neben der Etablierung der statistischen Auswertung der Datensätze werden auch die Methoden zum Testen des geeigneten Bildanalyse-Algorithmus und der Funktionalität der Mikroarrays beschrieben.

5.1.2 miRNA Mikroarray

Zur Herstellung des Mikroarrays wurde eine miRNA Bibliothek der Firma Ambion (www.ambion.com) (Shingara, Keiger et al. 2005) verwendet. Diese Bibliothek beinhaltet insgesamt 377 miRNAs, davon

- 312 humane miRNAs
- 51 murine miRNAs
- 42 Ratten miRNAs

- 4 positive Kontrollen
- 3 negative Kontrollen

Auf jedem Objektträger wird die gesamte Bibliothek drei Mal geprintet, so dass von jeder miRNA 3 technische Wiederholungsspots vorliegen (s. Abbildung 31).

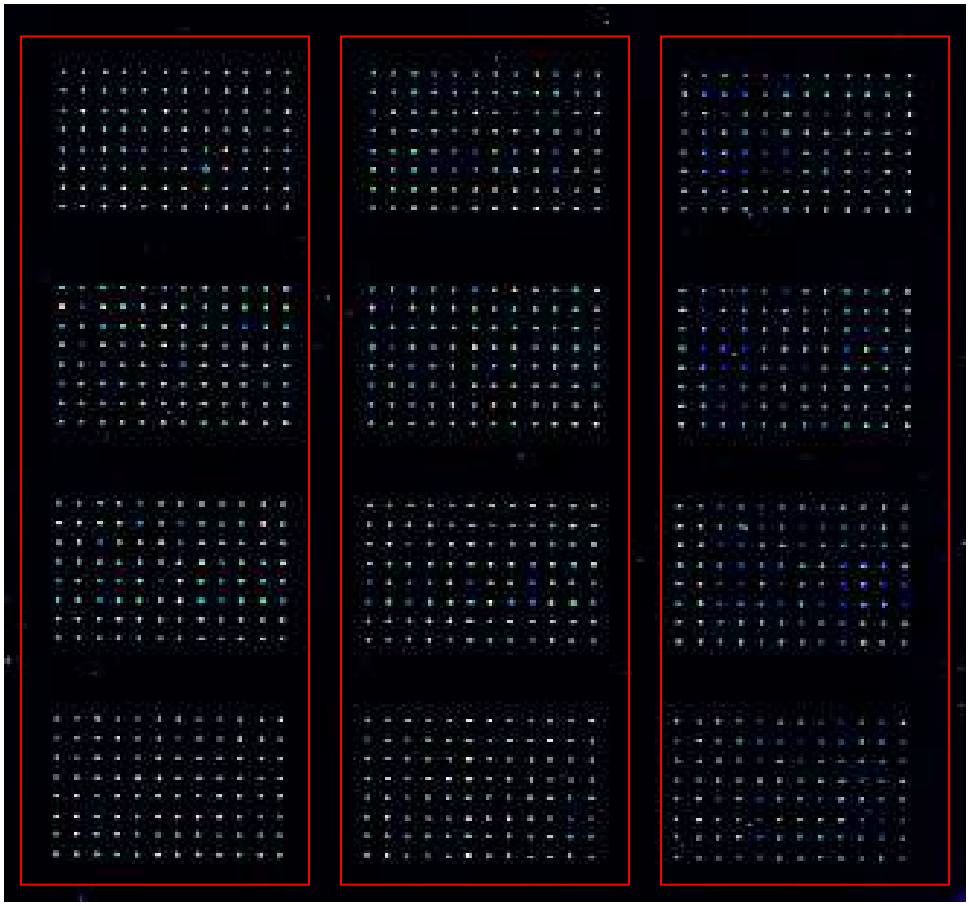


Abb. 31: Design des miRNA Mikroarray. Auf dem miRNA Mikroarray wird die gesamte Bibliothek 3 mal auf den Objektträger geprintet. Die einzelnen Subarrays sind durch rote Rahmen gekennzeichnet.

5.2 Vorversuch zum Testen der Bildanalyse-Algorithmen und der Funktionalität des miRNA Mikroarrays

Mit einem Vorversuch sollen zunächst die folgenden Fragen geklärt werden.

- Welches ist der beste Bildanalyse-Algorithmus zur Auswertung der miRNA-Mikroarrays?
- Können die miRNA-Mikroarrays reproduzierbar differentiell exprimierte miRNAs messen?

5.2.1 Datensatz

Um diese beiden Fragen zu überprüfen, werden zwei Proben mehrfach gegeneinander hybridisiert, Als Proben wurden SH-SY5Y und NB-K verwendet. SH-SY5Y ist eine humane Neuroblastom-Zelllinie, die von einem Tumor des Stadiums 4 abstammt (Pahlman, Odelstad et al. 1981). NB-K ist ein Neuroblastom-Primärtumor. Insgesamt werden die beiden Proben 5 mal gegeneinander hybridisiert.

5.2.2 Testen der Bildanalyse-Algorithmen zur Auswertung der Bilder der miRNA-Mikroarrays

Die Erwartung an einen geeigneten Bildanalyse-Algorithmus ist die exakte Berechnung der Spotintensitäten bzw. der daraus resultierenden $\log_2$-Verhältnisse aus den Pixelintensitäten des Spots unabhängig von der gegebenen Situation, d.h. der Spotgröße und -form, der Hintergrundintensität und anderen Parametern, die die Qualität der Spots beeinflussen können. Wie gut ein Bildanalyse-Algorithmus die Spotintensitäten berechnet, kann anhand der Reproduzierbarkeit von technischen Wiederholungsspots getestet werden. Da auf jedem Mikroarray 3 technische Wiederholungsspots für jede miRNA vorliegen, kann die Reproduzierbarkeit anhand dieser Spots überprüft werden. So sollten die 3 Wiederholungsspots der miRNAs gleiche Messwerte nach der Bildanalyse aufweisen. Es werden die drei verschiedenen Bildanalyse-Algorithmen, *Adaptive Threshold*, *Fixed circle* und *Histogram* getestet.

Die Daten werden zunächst nicht normalisiert, da wir Wiederholungsmessungen innerhalb eines Mikroarray betrachten, und somit Chip-abhängige systematische Effekte nicht relevant sind. Auch sollen die Wiederholungsspots der verschiedenen miRNAs ungefähr die gleiche Intensität haben, so dass eine Chip-interne intensitätsabhängige Normalisierung ebenfalls nicht notwendig ist. Es muss jedoch darauf geachtet werden, dass auf den Mikroarrays keine lokalen Effekte vorliegen, die die Intensität von einzelnen Wiederholungsspots beeinflussen.

Bildplots

In Abbildung 32 wird zunächst die Verteilung der $\log_2$-Ratios aller 5 Mikroarrays für die 3 verschiedenen Bildanalyse-Methoden im Bildplot betrachtet. Dies kann einen Hinweis auf mögliche lokale Effekte geben, die die Reproduzierbarkeit der Wiederholungsspots auf einem Array reduzieren können.

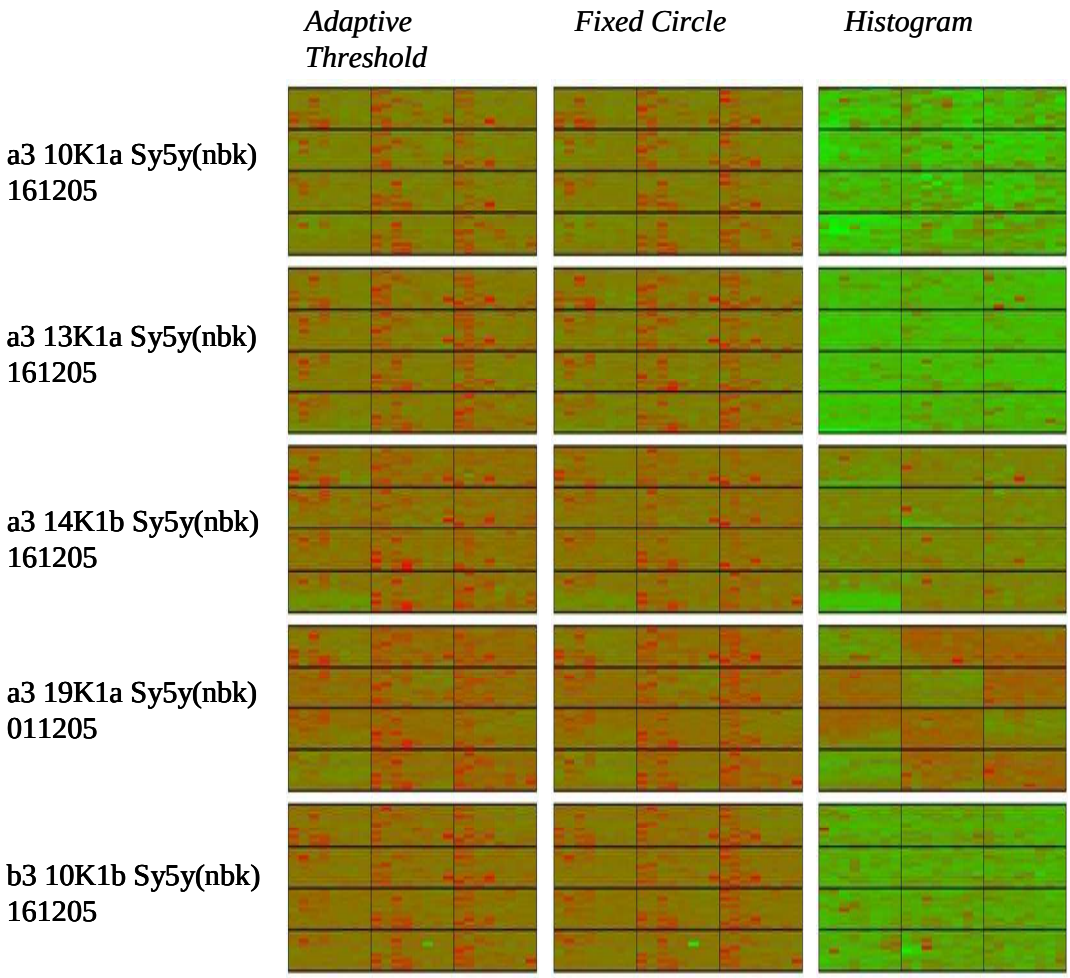


Abb. 32: Bildplots der 5 miRNA Mikroarrays des Vorversuches nach der Auswertung mit unterschiedlichen Bildanalyse-Algorithmen. Mit den Bildplots sollen die Mikroarrays auf lokale Effekte, die zur Reduzierung der Reproduzierbarkeit führen können, untersucht werden. Die Abbildung zeigt alle 5 miRNA-Mikroarrays des Vorversuches, die jeweils mit den Bildanalyse-Algorithmen *Adaptive Threshold*, *Fixed circle* und *Histogram* ausgewertet wurden. Während der *Adaptive Threshold*- und *Fixed circle*-Algorithmus zu Bildern mit zumeist gleichen Rot- bzw. Grünanteil führen, tritt bei dem *Histogram*-Algorithmus eine deutliche Verschiebung in den Grünbereich auf. Auf den Mikroarrays „a319K1aSy5y(nbk)011205“ und „a314K1b Sy5y(nbk)16120“ sind lokale Effekte zu sehen. Diese treten auch bei den anderen Bildanalyse-Algorithmen auf, hier jedoch wesentlich schwächer.

Die Mikroarrays a3 14K1b und a3 19K1a zeigen lokale Effekt auf, die vor allem bei der Auswertung mit der *Histogram*-Methode deutlich zu sehen sind, aber auch bei den anderen beiden Bildanalyse-Methoden –wenn auch schwächer– auftreten. Diese lokalen Effekte müssen bei der Bewertung der Reproduzierbarkeit der Wiederholungs-messungen berücksichtigt werden. Weiterhin ist auffällig, dass bei der *Histogram*-Auswertung die $\log_2$-Verhältnisse im Vergleich zu den anderen Algorithmen in Richtung des Cyanin-3 Kanals verschoben sind, was anhand des deutlich höheren grünen Anteil des Bildplots zu sehen ist.

Die Qualitätsparameter in den Tabellen 10-12 geben Aufschluss, ob der experimentelle Teil des Versuches, d.h. Aufreinigung der miRNA, Färbung und Hybrisierung, funktioniert hat, somit der Versuch zur Bewertung der Bildanalyse-Algorithmen geeignet ist. Zudem geben sie einen ersten Anhaltspunkt über die Reproduzierbarkeit der jeweiligen Subarrays. Die Qualitätsparameter sind in Tabelle 10 - 12 dargestellt. Diese Qualitätsparameter entsprechen den Parametern, die für die Auswertung von cDNA Mikroarrays verwendet werden und in Kapitel 2.2.6 beschrieben werden. Als zusätzlicher Parameter wird hier noch die mittlere Korrelation zwischen den 3 Subarrays berechnet.

Bei allen 3 Bildanalyse-Methoden und allen 5 Mikroarrays liegen die Signalintensitäten aller Spots über dem Hintergrund (s. Reihe „Anzahl der Gene über dem Hintergrund“). Dies zeigt, dass eine Hybridisierung auf den Mikroarrays erfolgt ist. Die hohe Korrelation zwischen den Wiederholungsspots (s. Reihe „Korrelation zwischen den Subarrays“), die bei den Bildanalyse-Algorithmen *Adaptive Threshold* und *Fixed Circle* zu sehen ist, deutet zusätzlich darauf hin, dass die Hybridisierung auch spezifisch erfolgt ist. Bei der Berechnung der Verhältnisse von Signal- zu Hintergrundintensität des Cyanin 3- und des Cyanin 5-Kanals, welche möglichst hohe Werte haben sollten, zeigt der Bildanalyse-Algorithmus *Fixed circle* die besten Verhältnisse. Für die Auswahl des geeigneten Bildanalyse Algorithmus ist jedoch eine gute Korrelation zwischen den Subarrays von noch größerer Bedeutung.

Die *Histogram*-Methode zeigt im Vergleich zu den beiden anderen Methoden deutlich schlechtere Qualitätsmerkmale. Anhand dieser Qualitätsparameter ist der Bildanalyse-Algorithmus *Fixed circle* für die Auswertung der Bilder am besten geeignet.

Tab. 10: Qualitätsparameter nach der Auswertung mit dem Bildanalyse-Algorithmus *Adaptive Threshold.*

	a3 10K1a Sy5y(nbk)	a3 13K1a Sy5y(nbk)	a3 14K1a Sy5y(nbk)	a3 19K1a Sy5y(nbk)	b3 10K1a Sy5y(nbk)
Fehlende Werte vor Hintergrundkorrektur	0	0	0	0	0
Fehlende Werte nach Hintergrundkorrektur	1	7	0	3	0
Anzahl der Gene über dem Hintergrund	1152	1150	1152	1152	1152
Verhältnis Signal/Hintergrund Cy3	3.2512	6.308	5.5279	4.52813	4.3829
Verhältnis Signal/Hintergrund Cy5	2.2475	2.7503	3.2889	2.5401	2.9011
Anzahl der Flags <100	0	0	0	0	0
Korrelation zwischen den Subarrays	0.52	0.87	0.62	0.58	0.57

Tab. 11: Qualitätsparameter nach der Auswertung mit dem Bildanalyse-Algorithmus *Fixed Circle*

	a3 10K1a Sy5y(nbk)	a3 13K1a Sy5y(nbk)	a3 14K1a Sy5y(nbk)	a3 19K1a Sy5y(nbk)	b3 10K1a Sy5y(nbk)
Fehlende Werte vor Hintergrundkorrektur	0	0	0	0	0
Fehlende Werte nach Hintergrundkorrektur	5	2	8	3	4
Anzahl der Gene über dem Hintergrund	1147	1150	1146	1149	1150
Verhältnis Signal/Hintergrund Cy3	4.4152	4.2009	9.1541	8.4725	6.4253
Verhältnis Signal/Hintergrund Cy5	2.7109	2.1911	4.8994	3.501	3.8034
Anzahl der Flags <100	0	0	0	0	0
Korrelation zwischen den Subarrays	0.68	0.75	0.76	0.79	0.88

Tab. 12: Qualitätsparameter nach der Auswertung mit dem Bildanalyse-Algorithmus *Histogram*.

	a3 10K1a Sy5y(nbk)	a3 13K1a Sy5y(nbk)	a3 14K1a Sy5y(nbk)	a3 19K1a Sy5y(nbk)	b3 10K1a Sy5y(nbk)
Fehlende Werte vor Hintergrundkorrektur	0	0	0	0	0
Fehlende Werte nach Hintergrundkorrektur	0	0	0	1	0
Anzahl der Gene über dem Hintergrund	1152	1152	1152	1152	1152
Verhältnis Signal/Hintergrund Cy3	1.8997	1.9335	2.7106	2.3796	2.2969
Verhältnis Signal/Hintergrund Cy5	2.1572	2.1918	2.3672	1.7723	2.5552
Anzahl der Flags <100	0	0	0	0	0
Korrelation zwischen den Subarrays	0.02	0.27	0.14	-0.42	0.12

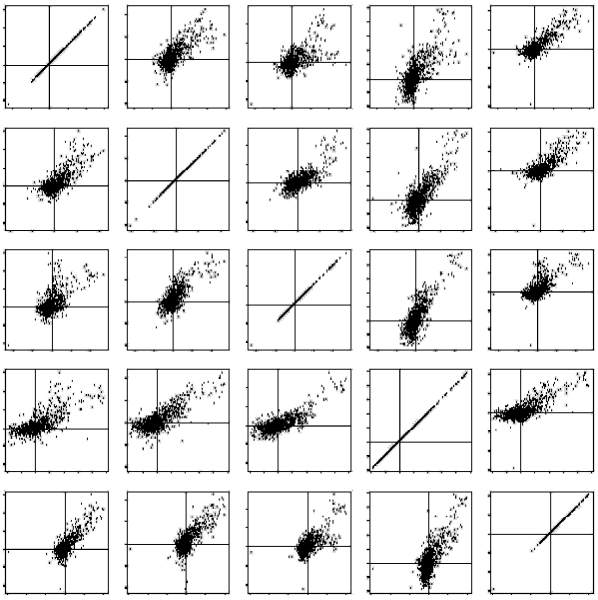
5.2.3 Können die miRNA-Mikroarrays reproduzierbar differentiell exprimierte miRNAs messen?

Mit der vorherigen Auswertung wurde gezeigt, dass die Subarrays eine deutliche Korrelation aufzeigen. Im nächsten Schritt soll nun geprüft werden, ob auch verschiedene miRNA Mikroarrays reproduzierbar die gleichen miRNAs als differentiell exprimiert messen. Hierzu wird eine Scatterplotmatrix bzw. eine Korrelationsmatrix der 5 Wiederholungs-miRNA-Mikroarrays erstellt. In der Scatterplotmatrix werden die $\log_2$-Verhältnisse der einzelnen Mikroarrays in einem xy-Koordinatensystem gegeneinander aufgetragen. Da wir hier verschiedene Mikroarrays miteinander vergleichen, muss zunächst eine Normalisierung der Daten durchgeführt werden. Da zu diesem Zeitpunkt noch kein „optimales“ Normalisierungsverfahren ausgewählt worden ist, werden zwei verschiedene Normalisierungen durchgeführt, die globale mediane Normalisierung und die Lowess-Normalisierung. Anhand beider Datensätze wird die Reproduzierbarkeit der Arrays beurteilt. Die Bewertung der Reproduzierbarkeit erfolgt

anhand der Darstellung von MM-Plots aller Mikroarrays gegeneinander bzw. der Berechnung der Korrelationskoeffizienten (s. Abbildung 33 und 34).

Bei beiden Normalisierungstechniken tritt eine deutliche Korrelation zwischen den fünf miRNA-Mikroarrays auf. Hiermit kann die Funktionalität der Mikroarrays gezeigt werden, d.h., es können differentiell exprimierte miRNAs reproduzierbar gemessen werden.

a)

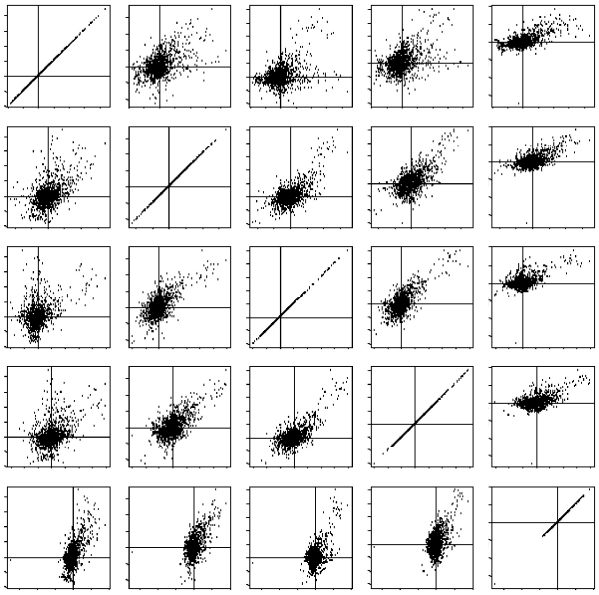


b)

	a3 10K1a Sy5y(nbk)	a3 13K1a Sy5y(nbk)	a3 14K1a Sy5y(nbk)	a3 19K1a Sy5y(nbk)	b3 10K1a Sy5y(nbk)
a3 10K1a Sy5y(nbk)	1	0.74	0.6	0.7	0.8
a3 13K1a Sy5y(nbk)	0.74	1	0.73	0.76	0.76
a3 14K1a Sy5y(nbk)	0.6	0.73	1	0.79	9.67
a3 19K1a Sy5y(nbk)	0.7	0.76	0.79	1	0.76
b3 10K1a Sy5y(nbk)	0.8	0.76	0.67	0.76	1

Abb. 33: MM-Plot und Korrelationsmatrix der log2 Verhältnisse der 5 Wiederholungs-Mikroarrays nach globaler medianer Normalisierung. (a) Die Abbildung zeigt einen MM-Plot der 5 Wiederholung-Mikroarrays des Vorversuches, nachdem eine globale mediane Normalisierung durchgeführt wurde. In dem MM-Plot werden die $\log_2$-Verhältnisse der einzelnen Mikroarrays jeweils gegeneinander im Koordinatensystem aufgetragen. Hiermit soll dargestellt werden, wie hoch die Reproduzierbarkeit der Daten ist. Die Mikroarrays zeigen eine gute Reproduzierbarkeit, was anhand der Anordnung der Datenpunkte entlang einer „Achse“ die von links unten nach rechts oben verläuft zu sehen ist. Diese Reproduzierbarkeit wird durch die Korrelationsmatrix in (b) bestätigt, die gute Korrelationskoeffizienten für alle Kombinationen aufweist. Der Korrelationskoeffizient ist ein Maß für den linearen Zusammenhang von zwei intervallskalierten Merkmalen. Er kann Werte zwischen -1 und 1 annehmen, Bei einem Wert von 1 besteht ein vollständig positiver bzw. bei einem Wert von -1 ein vollständig negativer linearer Zusammenhang. Bei einem Korrelationskoeffizient von 0 besteht kein Zusammenhang zwischen den Variablen.

a)



b)

	a3 10K1a Sy5y(nbk)	a3 13K1a Sy5y(nbk)	a3 14K1a Sy5y(nbk)	a3 19K1a Sy5y(nbk)	b3 10K1a Sy5y(nbk)
a3 10K1a Sy5y(nbk)	1	0.5088	0.3757	0.4299	0.6308
a3 13K1a Sy5y(nbk)	0.5088	1	0.657	0.6473	0.6092
a3 14K1a Sy5y(nbk)	0.3757	0.657	1	0.7275	0.5502
a3 19K1a Sy5y(nbk)	0.4299	0.6473	0.7275	1	0.5273
b3 10K1a Sy5y(nbk)	0.6308	0.6092	0.5502	0.5273	1

Abb. 34: Scatterplotmatrix (a) und Korrelationsmatrix (b) der log2 Verhältnisse der 5 Wiederholungs-Mikroarrays nach der Lowess-Normalisierung. Die Abbildung (a) zeigt einen MM-Plot der 5 Wiederholung-Mikroarrays des Vorversuches, nachdem eine Lowess- Normalisierung durchgeführt wurde. In dem MM-Plot werden die log2-Verhältnisse der einzelnen Mikroarrays jeweils gegeneinander im Koordinatensystem aufgetragen. Hiermit soll getestet werden, wie hoch die Reproduzierbarkeit der Daten ist. Die Mikroarrays zeigen eine gute Reproduzierbarkeit, die jedoch geringer ist, als die Reproduzierbarkeit nach einer globalen Normalisierung. Dies ist sowohl in den MM-Plots (a) als auch in der Korrelationsmatrix (b) zu sehen.

5.3 Untersuchung der Rolle von N-Myc auf die miRNA Expression *in vitro* und *in vivo*

Die miRNA Mikroarray-Plattform wird dann in einem Experiment verwendet, bei dem die Auswirkung des N-Myc Status auf die miRNA Expression *in vivo* und *in vitro* untersucht werden soll.

5.3.1 Der Transkriptionsfaktor N-Myc und seine Bedeutung im Neuroblastom

N-Myc ist ein Transkriptionsfaktor, der eine wichtige Rolle als prognostischer Faktor im Neuroblastom spielt. Das Neuroblastom ist ein Tumor, der aus Neuronen des sympathischen Nervensystems hervorgeht und entlang des sympathischen Grenzstrangs oder im Nebennierenmark lokalisiert ist. Er macht ungefähr 8 % aller Krebserkrankungen im Kindes- und Jugendalter aus und ist damit nach den Hirntumoren der häufigste solide Tumor im Kindesalter. Der Verlauf beim Neuroblastom kann sehr unterschiedlich sein und reicht von der spontanen Regression über die fortschreitende Erkrankung bis hin zur Metastasenbildung. Eine erhöhte N-Myc Expression als Folge der Amplifizierung des Gens erhöht die Aggressivität des Tumors (Schwab, Alitalo et al. 1983; Brodeur, Seeger et al. 1984).

5.3.2 Untersuchung der Rolle von N-Myc auf die miRNA Expression *in vivo*

Datensatz

Zur Untersuchung des N-Myc-Effektes auf die miRNA Expression *in vivo* werden 16 Neuroblastom-Proben ohne N-Myc Amplifikation mit 8 Neuroblastom-Proben mit N-Myc- Amplifikation verglichen. Als experimentelles Design wurde das Referenzdesign gewählt.

Auswertung des Datensatzes

Die Mikroarrays werden, entsprechend der Ergebnisse im Vorversuch, mit dem Bildanalyse-Algorithmus *Fixed Circle* ausgewertet.
Es werden verschiedene in Kapitel 2.2.5 beschriebene Qualitätsparameter erhoben. Alle Mikroarrays weisen Qualitätsparameter auf, die zulassen, dass sie mit in die weitere Auswertung eingehen.

Auswahl eines geeigneten Normalisierungsverfahrens

Es gibt eine Reihe von verschiedenen Faktoren, die innerhalb von Mikroarray-Versuchen zu systematischen Fehlern führen können. Um diese zu beseitigen und damit auch die biologischen Effekte verschiedene Mikroarrays miteinander vergleichbar machen zu können, müssen die Daten normalisiert werden. Hierfür gibt es verschiedene

Methoden, die in Kapitel 2.2.7 vorgestellt werden. Je nach Methode werden unterschiedliche Effekte korrigiert. So korrigiert z.B. die Lowess-Normalisierung intensitätsabhängige Abweichungen. Um sich für die richtige Methode entscheiden zu können, muss dementsprechend zunächst geprüft werden, welche Effekte zu korrigieren sind. Zudem ist zu beachten, dass jede dieser Methoden bestimmte Annahmen voraus setzt, die gegeben sein müssen, um die Methode effizient anwenden zu können. So muss z.B. für die Lowess-Normalisierung eine ausreichende Anzahl an Datenpunkten vorhanden sein, die die Abschätzung der für die Normalisierung notwendigen intensitätsabhängigen Regressionsgeraden ermöglicht. Um für diesen Versuch die geeignete Normalisierungsmethode auszuwählen, wird im ersten Schritt für die einzelnen Mikroarrays anhand der MA-Plots überprüft, ob intensitätsabhängige Effekte vorliegen. Ein Teil der Mikroarrays weisen intensitätsabhängige Effekte auf, zwei Beispiele sind in Abbildung 35 dargestellt.

Um diese zu korrigieren, wäre die Lowess-Normalisierung sinnvoll. Da es sich hier aber um eine Mikroarray-Plattform handelt, die mit 1152 Spots in jedem Subarray sehr klein ist, wird die Voraussetzung, dass ausreichend Datenpunkte zum Abschätzen der intensitätsabhängigen Regressionsgeraden vorhanden sein sollen, nicht erfüllt. Welche Auswirkungen die Lowess-Normalisierung auf die Verteilung des Datensatzes hat, ist in Abbildung 35 dargestellt. Die Normalisierung führt z.B. bei den Mikroarray ae und ag zur Erhöhung der Streuung vor allem in höheren Intensitätsbereichen, da vor allem in diesem Bereich nur sehr wenige Datenpunkte vorhanden sind, die keine genaue Abschätzung der Regressionsgerade erlauben.

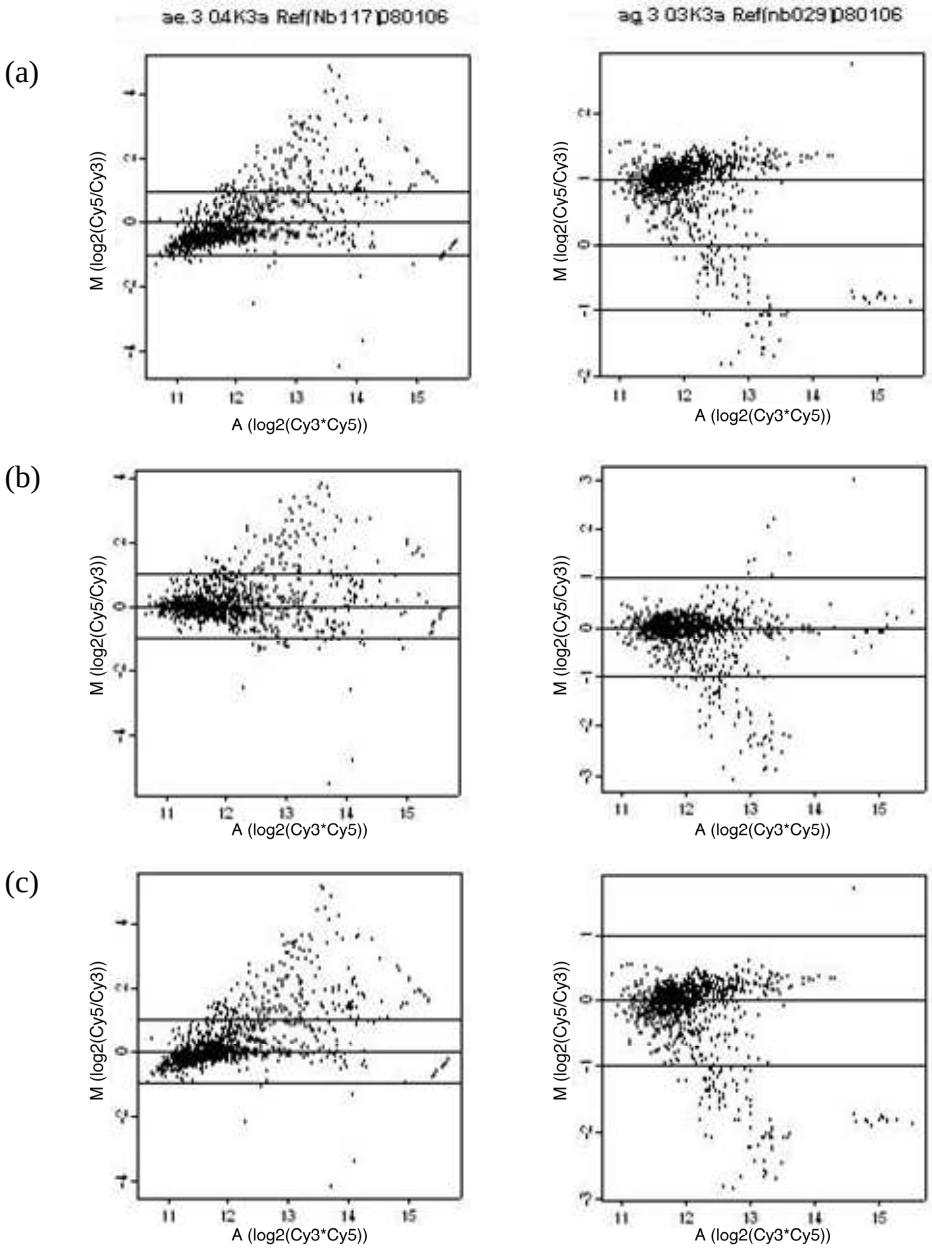


Abb. 35: MA-Plots der Mikroarrays *ae 304K3a Ref(Nb117)080106* und *ag 303K3a Ref(nb029)080106* vor der Normalisierung (a) nach der Lowess-Normalisierung (b) und nach der globalen medianen Normalisierung (c). Der MA-Plot stellt auf der x-Achse die mittlere $\log_2$-Intensität der beiden Kanäle jedes Spots als A-Wert und auf der y-Achse das Verhältnis der $\log_2$ Intensitäten der beiden Kanäle als M-Wert dar. (a) Beide Mikroarrays zeigen intensitätsabhängige Effekte, d.h. mit zunehmender Gesamtintensität des Spots (A-Wert) verschiebt sich die „Punktwolke" bei dem Mikroarray in der linken Abbildung in den positiven bzw. bei dem Mikroarray in der rechten Abbildung in den negativen Bereich. (b) Die Lowess-Normalisierung korrigiert die intensitätsabhängigen Effekte. Aufgrund der geringen Anzahl an Spots, welche zur Abschätzung der Regressionsgerade zur Verfügung stehen, erfolgt möglicherweise eine „Überkorrektur", d.h. die $\log_2$-Verhältnisse differentiell exprimierte miRNA werden in Richtung 0 korrigiert. (c) Bei der globalen medianen Normalisierung erfolgt eine Korrektur der M-Werte, so dass das mittlere log2-Verhältnis über alle Gene 0 ist. Die intensitätsabhängigen Effekt bleiben hierbei bestehen.

Als Alternative zur Lowess-Normalisierung wird eine globale mediane Normalisierung durchgeführt, deren Ergebnis als MA-Plots in der zweiten Spalte von Abbildung 35 dargestellt ist. Die intensitätsabhängigen Effekte werden mit dieser Methode nicht korrigiert.

Auch eine Normalisierung über ein Subset von Genen ist nicht möglich, da nur wenige Kontrollen auf dem Chip sind, die zudem nicht auf verschiedene Intensitätsbereiche verteilt sind. Eine eindeutige Auswahl der zu verwendenden Normalisierungstechnik ist anhand der gegebenen Situation zunächst nicht möglich.

Ziel der Normalisierung ist die Reduktion von systematischen Veränderungen und die Erhöhung der biologischen Reproduzierbarkeit. Um zu testen, welche Normalisierungstechnik unter den gegebenen Bedingungen dennoch am besten geeignet ist, werden die Daten mit den verschiedenen Normalisierungsmethoden korrigiert und im Anschluss mit einer Clusteranalyse (*average linkage clustering*, euklidischer Abstand) geprüft, bei welcher der Methoden die Proben mit dem gleichen Phänotyp ein gemeinsames Cluster bilden. Neben der globalen medianen Normalisierung und der Lowess-Normalisierung mit den Standardeinstellungen, die bei der cDNA-Mikroarray-Auswertung verwendet werden, werden zusätzlich weitere Einstellungen für die Lowess-Normalisierung getestet.

Als Parameter für die Lowess-Normalisierung können die „Spanne“, die den Anteil an Spots, die zur Berechung einer lokalen Regressionsgerade verwendet werden, und die „Iteration“, die die Anzahl der Schätzung der Regressionsgeraden mit jeweils neuen Robustheitsgewichten angibt, variiert werden. Durch die Erhöhung der „Spanne“ werden mehr miRNA-Datenpunkte in die Abschätzung der Regressionsgerade mit einbezogen, bei der Erhöhung der Anzahl der Iterationen wird der Algorithmus robuster gegenüber Ausreißern.

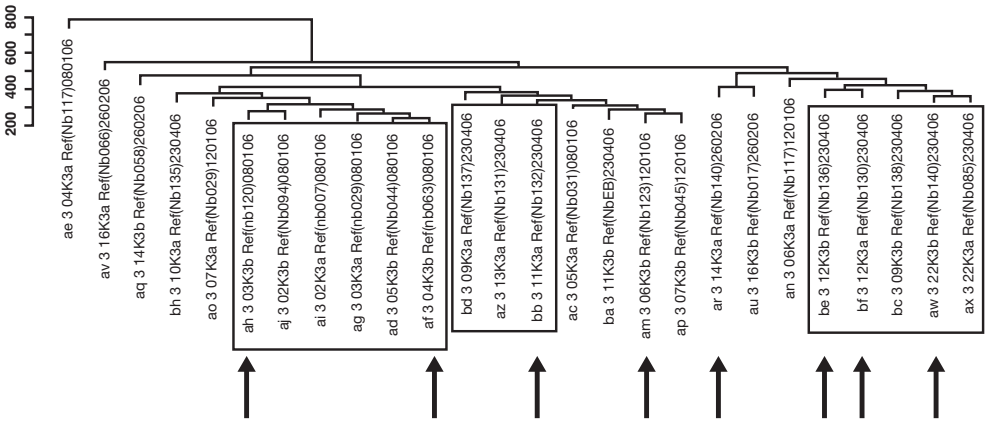
Folgende Einstellungen werden getestet:

Lowess-Normalisierung	Spanne:0.5	Iteration:3
Lowess-Normalisierung	Spanne:0.5	Iteration:10
Lowess-Normalisierung	Spanne:3	Iteration:3
Lowess-Normalisierung	Spanne:3	Iteration:10

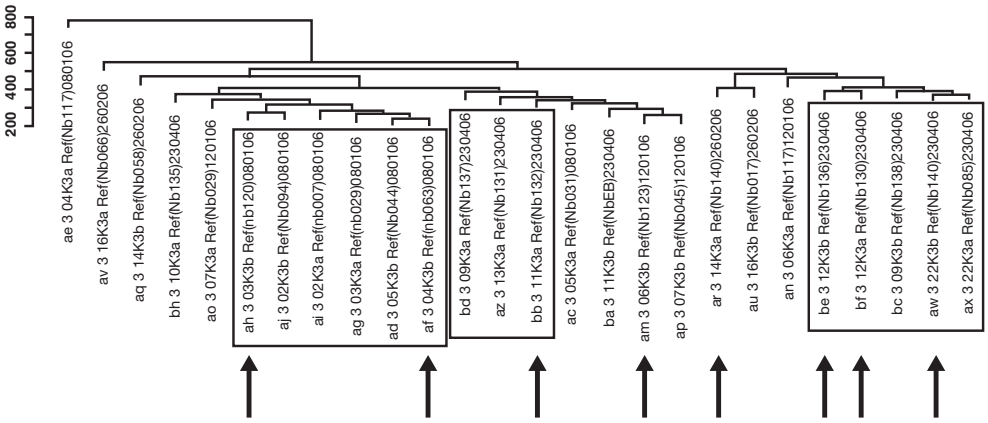
globale mediane Normalisierung

In Abbildung 36 sind die Dendrogramme bei der Verwendung der verschiedenen Einstellungen dargestellt

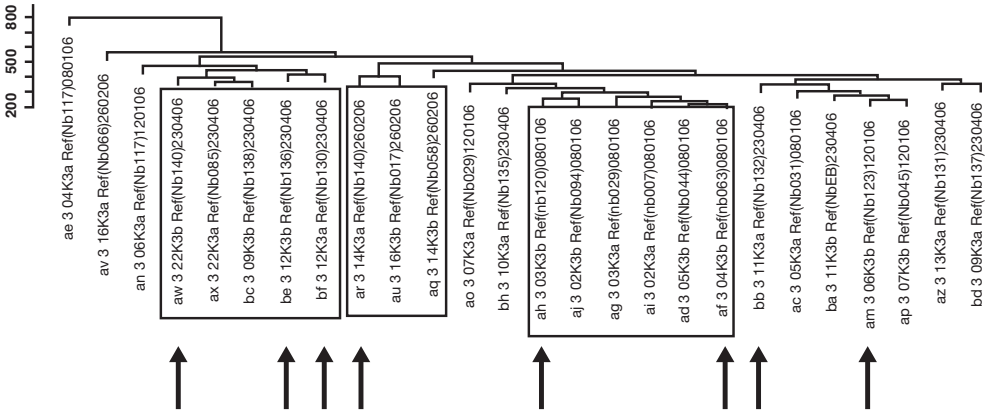
a) Lowess-Normalisierung Span:0.5 Iteration:3



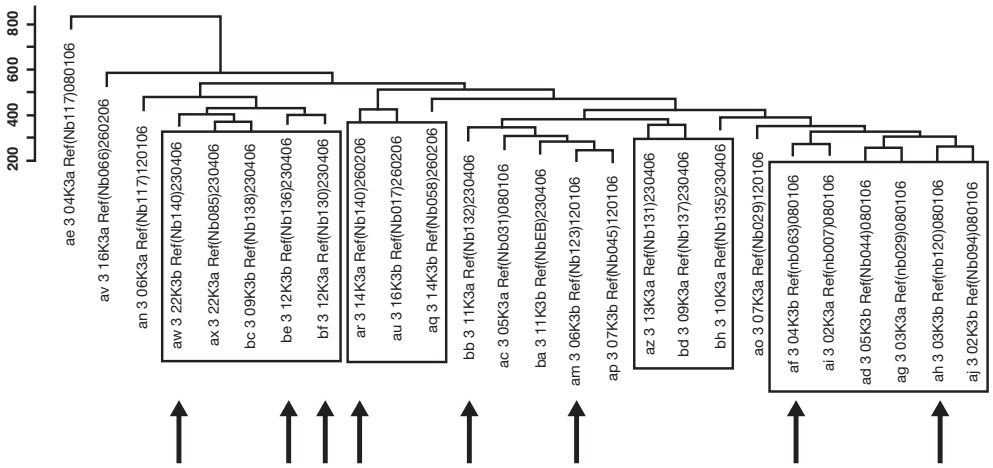
b) Lowess-Normalisierung Span:0.5 Iteration:10



c) Lowess-Normalisierung Span:3 Iteration:3



d) Lowess-Normalisierung Span:3 Iteration:10



e) globale mediane Normalisierung

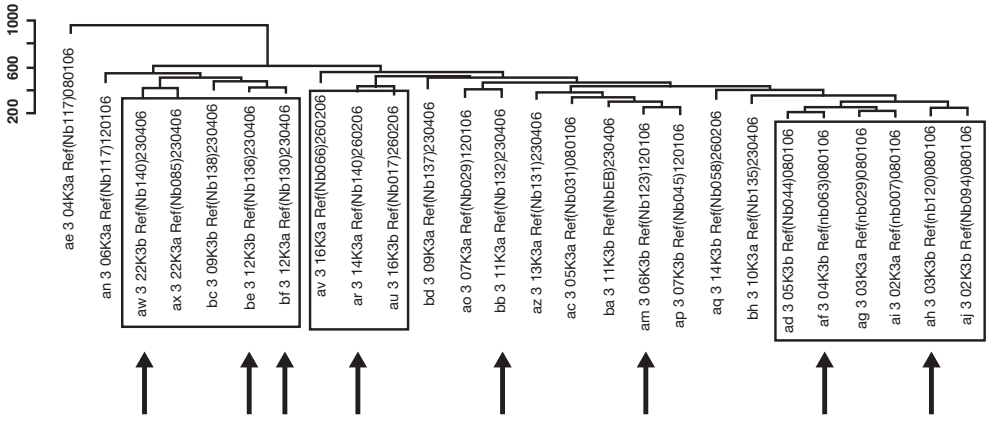


Abb. 36: Cluster-Dendogramme (Euklidische Distanz, hierarchisches Clustern mit durchschnittlichem Abstand) zur Darstellung der Ähnlichkeiten der Ergebnisse nach der Verwendung von unterschiedlichen Parametern bei der Lowess-Normalisierung bzw. der globalen medianen Normalisierung. Das Clusterverfahren ist ein unüberwachtes Analyseverfahren, mit dem die miRNA-Mikroarrays entsprechend ihrer Ähnlichkeiten bzw. Unterschiede in Gruppen zusammenclustern. Je größer die Ähnlichkeiten der Proben, desto näher liegen die Mikorarrays innerhalb der Cluster zusammen. Da das Ziel der Normalisierung die Eliminierung von technischen Varianzen zur besseren Fokussierung auf die biologischen Unterschiede ist, sollte nach einer guten Normalisierung biologische Wiederholungsmessungen ein gemeinsames Cluster bilden. Anhand dieser Clusteranalyse soll überprüft werden, welches der Normalisierungsmethoden am Besten geeignet ist. Bei allen hier verwendeten Normalsierungsmethoden bilden sich Cluster von miRNA-Mikroarrays mit Neuroblastomproben, die am gleichen Tag hybridisiert wurden. Diese sind in der Abbildung mit einem Kasten umrandet. Proben die eine N-Myc Amplifikation haben, sind durch einen Pfeil markiert.

Keine der verschiedenen Normalisierungsmethoden führt zu einer Gruppierung der Tumor-Subtypen N-Myc-amplifiziert und N-Myc-Nicht-amplifiziert in verschiedene Cluster. Dagegen treten immer wieder Subcluster auf, die einem bestimmten Hybridisierungsdatum zugeordnet werden können. Diese Cluster sind in den Abbildungen durch Kästen markiert. Die experimentell bedingten technischen

Unterschiede scheinen somit größer zu sein als die biologischen Unterschiede, welche in Abhängigkeit von der N-Myc-Amplifizierung auftreten.
Keine der Normalisierungsmethoden ist somit gut geeignet. Um technische Unterschiede dennoch zu reduzieren und damit die Zahl der falsch positiven miRNAs zu verringern wurde die globale mediane Normalisierungsmethode angewendet. Diese Methode wurde ausgewählt, da die alternative Lowess-Normalisierung vor allem die Datenpunkte im höheren Intensitätsbereich möglicherweise „überkorrigiert" – ein Intensitätsbereich, in dem die differentiell exprimierten miRNAs voraussichtlich liegen. Dies würde dazu führen, dass differentiell exprimierte miRNAs in diesem Bereich nicht detektiert werden.

Selektion differentiell exprimierter miRNAs

Die Berechnung differentiell exprimierter miRNAs erfolgt mit der SAM-Statistik (s. Kapitel 2.2.7). Werden alle 384 miRNAs in die Auswertung mit einbezogen, ist keine der miRNAs signifikant differentiell exprimiert. Dies wird u.a. bedingt durch die Korrektur für das multiple Testproblem. Um die Anzahl der durchzuführenden Tests zu reduzieren, wird der Datensatz basierend auf der Standardabweichung über alle Mikroarrays auf 154 miRNAs reduziert. Dies ist erlaubt, wenn anhand eines Parameters gefiltert wird, der nicht im Zusammenhang mit den zu vergleichenden Phänotypen steht. Durch den hier verwendeten Filter werden miRNAs aus der Auswertung ausgeschlossen, die aufgrund der geringen Intensitätsunterschiede zwischen allen Arrays auch keine Unterschiede zwischen den beiden Phänotypen ausgewiesen hätten.

Insgesamt sind, bei einer *false discovery rate* (FDR) von <5%, 17 miRNAs in den N-Myc amplifizierten Proben signifikant hoch- und keine der miRNAs herunterreguliert. Die Liste der differentiell exprimierten miRNAs ist in Tabelle 13 dargestellt.

Innerhalb der Ergebnisliste finden sich 2 der positiven Kontrollen. Da diese Kontrolle in beiden Proben in gleicher Menge vorliegt, hätten wir für beide ein $\log_2$-Verhältnis von 0 erwartet. Der MA-Plot des Gesamtergebnisses zeigt eine Verschiebung des gesamten Plots im höheren Intensitätsbereich und damit in den positiven log2 Bereich. Hiermit zeigen sich die intensitätsabhängigen Effekte, die durch die globale mediane Normalisierung nicht korrigiert wurden.

Tab. 13: Liste der miRNAs, die im Vergleich der N-Myc amplifizierten Neuroblastomproben mit den Neuroblastomproben, die nur eine N-Myc Kopie haben signifikant differentiell exprimierten sind

miRNA	log2 Verhältnis N-Myc Amp/ N-Myc NonAmp	Richtung N-Myc Amp/ N-Myc NonAmp	FDR q-value(%)
hsa_miR_199a	0.9	up	0
hsa_miR_199a_AS	1.04	up	0
hsa_miR_20	1.15	up	0
hsa_miR_25	0.59	up	0
hsa_miR_17_5p	1.08	up	0
hsa_miR_106a	1.2	up	0
hsa_miR_18	0.84	up	0
hsa_miR_181a	0.84	up	0
hsa_miR_181b	0.8	up	0
hsa_miR_221	0.43	up	0
hsa_miR_92	1.05	up	0
hsa_miR_93	0.8	up	0
hsa_miR_99a	0.69	up	0
mmu_miR_106a	1.06	up	0
Control_1	0.71	up	0
hsa_let_7b	0.75	up	0
Control_1	0.71	up	0

In Abbildung 37 wird der MA-Plot des Gesamtergebnisses dargestellt.

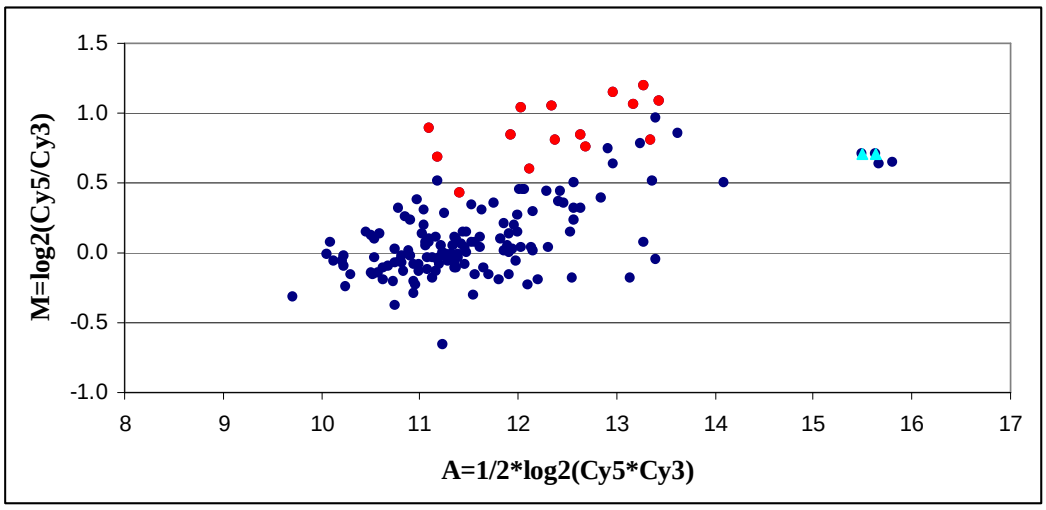


Abb. 37: MA-Plot der differentiellen miRNA Expression von N-Myc amplifiziert und N-Myc nicht-amplifiziert Neuroblastomproben. Der MA-Plot stellt für jeden Spot die mittlere $\log_2$-Intensität der beiden Kanäle (A-Wert) gegen das Verhältnis der $\log_2$ Intensitäten (M-Wert) dar. Mit den blauen Spots ist der Gesamtdatensatz dargestellt, die roten Spots sind die signifikant differentiell regulierten miRNAs, die hellblauen Spots die positiven Kontrollen.

5.3.3 Untersuchung der Rolle von N-Myc auf die miRNA Expression *in vitro*

In einem zweiten Versuch wurde die Wirkung des N-Myc-Levels auf das miRNA-Profil *in vitro* untersucht.

Datensatz

Die Neuroblastomzelllinie SH-EP, die keine N-Myc Amplifikation hat, wurde mit einer cDNA transfiziert, die ein Fusionsprotein aus N-Myc und einer Estrogen-responsiven (ER) Domäne kodiert. Die ER-Domäne ist mutiert, so dass nur 4-Hydroxytamoxifen (4-OHT) bindet und nicht das Wildtyp Estrogen (Eilers, Picard et al. 1989; Littlewood, Hancock et al. 1995). Dieses System ermöglicht eine genaue Aktivierung der N-Myc Expression mittels der Zugabe von 4-Hydroxytamoxifen (4-OHT).

Auswertung der Datensatzes

Da nur zwei Proben miteinander verglichen werden, wird hier das Paardesign gewählt, bei dem die zu vergleichenden Proben gegeneinander auf dem Mikroarray hybridisiert werden. Um die Varianz der Daten, die aufgrund des unterschiedlichen Verhaltens der Farbstoffe auftritt, zu reduzieren, werden zwei Wiederholungsversuche mit getauschten Farbstoffen – ein sogenanntes *Dye swap* Experiment – durchgeführt. Für jede Probe wurden zwei miRNA Mikroarrays mit jeweils 3 Subarrays hybridisiert.

Abb. 38: Experimentelles Design zur Untersuchung der N-Myc abhängigen miRNA Expression in vitro. Es wurde ein Dye-Swap-Experiment durchgeführt (s. Kapitel 2.2.2)

Auch hier werden Qualitätskontrollen entsprechend der in Kapitel 2.2.5 beschriebenen Methoden für jeden Mikroarray gemacht. Alle Mikroarrays zeigen gute Qualitätsparameter, so dass sie in die weitere Auswertung mit eingehen.

Da intensitätsabhängige Effekte nur im Niedrigintensitätsbereich auftreten (s. Abbildung 39), ein Bereich der stark vom Hintergrund beeinflusst wird und damit sowieso keine Aussagen über mögliche differentiell exprimierte miRNAs zulässt, wird eine globale mediane Normalisierung verwendet.

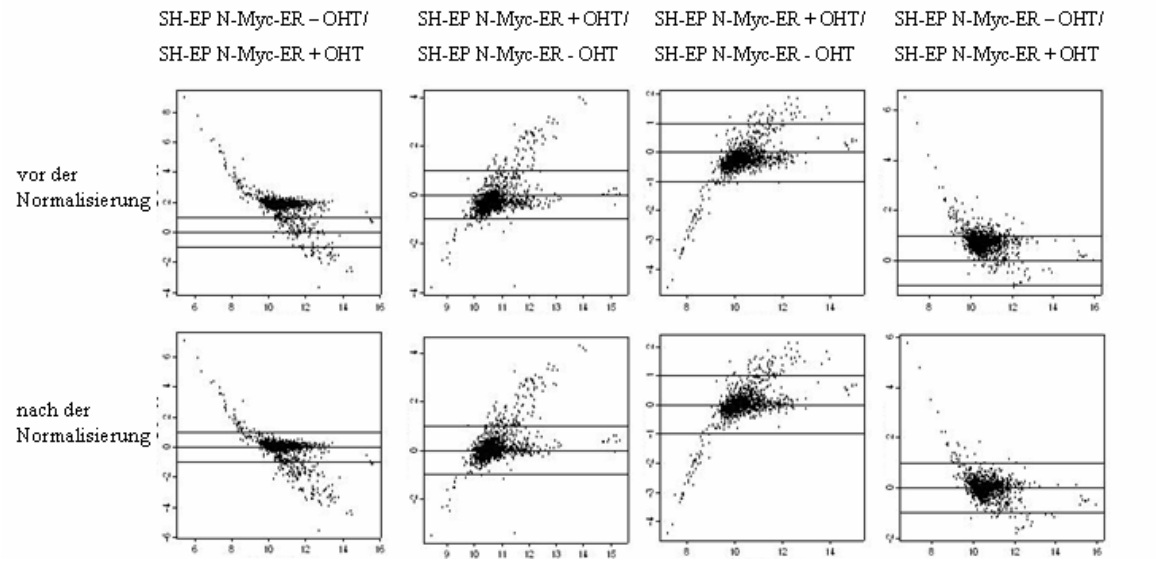


Abb. 39: MA-Plots der 4 Mikroarrays vor und nach der globalen medianen Normalisierung. Der MA-Plot stellt auf der x-Achse die mittlere $\log_2$-Intensität der beiden Kanäle eines jeden Spots als A-Wert und auf der y-Achse das Verhältnis der $\log_2$ Intensitäten der beiden Kanäle als M-Wert dar. In der 1. Reihe sind die MA-Plots vor und in der 2. Reihe nach der Normalisierung. Durch die Normalisierung erfolgt eine Verschiebung der „Punktwolke" auf die x-Achse, d.h. das mittlere $\log_2$-Verhältnisse beträgt Null.

Die Selektion von differentiell exprimierten miRNAs erfolgt anhand der „*Volcanoplot*-Methode" (s.2.2.7). Hierzu wird für jede miRNA das $\log_2$-Verhältnis auf der x-Achse gegen den absoluten Wert der t-Statistik auf der y-Achse aufgetragen. miRNAs, die einem absoluten $\log_2$ Ratio von größer als 1 und einer absoluten t-Wert von größer als 1,96 haben, werden als differentiell exprimiert selektiert.

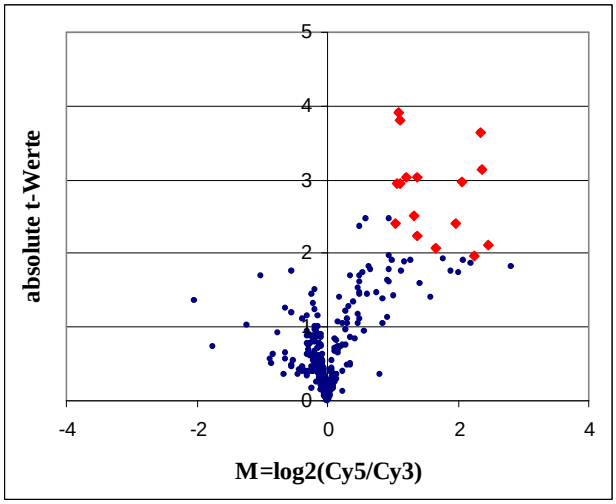


Abb. 40: *Volcanoplot* der Auswertung SHEP-Myc-ER + 4-OHT gegen SHEP-Myc-ER. Im *Volcanoplot* wird das Verhältnis der $\log_2$ Intensitäten der beiden Kanäle als M-Wert auf der x-Achse gegen den absoluten t-Wert auf der y-Achse aufgetragen Differentiell exprimiert selektierte miRNAs sind als rote Punkte markiert

Insgesamt sind 16 miRNAs differentiell exprimiert, die in der folgenden Tabelle aufgelistet sind.

Tab. 14: Differentiell exprimierte miRNAs des *in vitro* Vergleiches SHEP-Myc-ER + 4-OHT gegen SHEP-Myc-ER.

Name	log2 Verhältnis NMyc-ER + OHT/ NMyc-ER	Richtung NMyc-ER + OHT/ NMyc-ER	absolute t-Wert
hsa_miR_92	1.09	up	3.91
mmu_miR_106a	1.1	up	3.8
hsa_let_7c	2.33	up	3.64
hsa_let_7b	2.38	up	3.14
hsa_let_7f	1.37	up	3.04
hsa_miR_17_5p	1.21	up	3.03
hsa_let_7d	2.07	up	2.97
hsa_miR_320	1.07	up	2.95
hsa_miR_93	1.11	up	2.94
hsa_miR_106a	1.33	up	2.51
hsa_miR_100	1.95	up	2.41
hsa_miR_30d	1.03	up	2.4
hsa_miR_99a	1.36	up	2.23
hsa_miR_31	2.47	up	2.11
hsa_miR_145	1.66	up	2.07
hsa_miR_221	2.25	up	1.96

5.3.4 Vergleich der N-Myc abhängigen miRNA Expression *in vivo* und *in vitro*

Vergleicht man die Ergebnisse der N-Myc abhängigen miRNA Expression *in vivo* mit denen des *in vitro* Versuches, zeigt sich eine Überschneidung von 8 miRNAs.

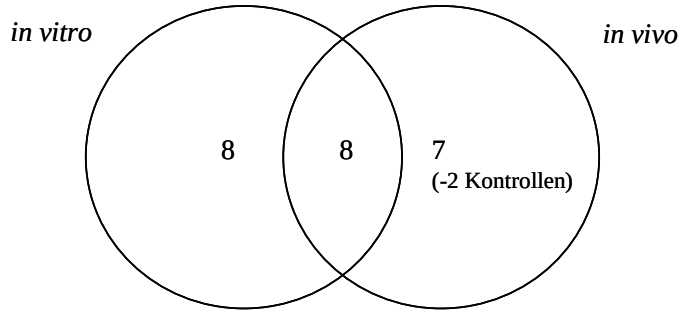


Abb. 41: Venndiagramm der Überschneidungen zwischen den Ergebnissen der N-Myc induzierten miRNA Expression *in vivo* und *in vitro*.

Folgende miRNAs treten in beiden Ergebnislisten auf:

hsa_miR_99a
hsa_miR_93
hsa_miR_92
mmu_miR_106a
hsa_miR_221
hsa_miR_17_5p
hsa_miR_106a
hsa_let_7b
(Schulte, Horn et al. 2008)

5.4 Diskussion

In diesem Kapitel wird die Etablierung der miRNA-Mikroarraytechnologie zur simultanen Messung der Genexpression von 377 miRNAs beschrieben. Mit den Mikroarrays wurden zunächst Vorversuche durchgeführt, mit denen ein geeigneter Bildanalyse-Algorithmus bzw. die Funktionalität des Mikroarrays, d.h. die Fähigkeit differentiell exprimierte miRNAs reproduzierbar messen zu können, getestet wurden. Nachdem die „Funktionalität“ der Plattform gezeigt worden ist, wurden mit der Plattform zwei Versuche durchgeführt, in denen die Auswirkung von N-Myc auf die miRNA Expression sowohl *in vitro* als auch *in vivo* untersucht wurde.

Die Vorversuche haben gezeigt, dass mit dem Bildanalyse-Algorithmus *Fixed circle* die größte Reproduzierbarkeit zwischen den Wiederholungsspots auf einem Array erzielt werden kann. Die Intensität der Spots konnte unabhängig von störenden Parametern, wie die Spotgröße und –form, mittels des Algorithmus gut dargestellt werden. Weiterhin konnte im Vorversuch gezeigt werden, dass verschiedene Mikroarrays die miRNA Expressionslevel reproduzierbar messen können. Neben der *Fixed circle* Methode zeigt auch der *Adaptive Threshold* Algorithmus gute Ergebnisse, dagegen ist die Reproduzierbarkeit beim *Histogram* Algorithmus deutlich schlechter. Das bei dieser Methode auftretende schlechte Verhältnis zwischen den Signal- und den Hintergrundintensitäten lässt annehmen, dass bei der Segmentierung keine gute Trennung zwischen Signal und Hintergrund erfolgt ist. Hiermit ließe sich auch die Verschiebung der $\log_2$-Verhältnisse in den grünen Bereich erklären, da ein höherer Hintergrund im roten Kanal die Signalintensitäten in diesem Kanal reduziert und damit eine solche Verschiebung verursachen würde. Eine Veränderung des Rasters mit einer geringeren Spotgröße könnte möglicherweise zu einem besseren Ergebnissen des Algorithmus führen.

Bei der Normalisierung der Datensätze konnte aufgrund des vorliegenden Designs der miRNA-Mikroarray-Plattform, d.h. geringe Anzahl von Spots bzw. fehlende Kontrollen, keine gut geeignete Methode verwendet werden. Um geeignete Normalisierungsmethoden verwenden zu können, was zu einer Reduzierung der falsch positiven und negativen Ergebnisse führen würde, müsste eine der folgenden Veränderungen der Chip-Struktur vorgenommen werden. Entweder müsste die Gesamtzahl der Spots auf den Arrays erhöht werden oder aber eine ausreichende Anzahl von Kontrollen in unterschiedlichen Intensitätsbereichen zusätzlich auf den Mikroarray geprintet werden, aus denen dann eine Regressionsgerade für die Lowess-Normalisierung abgeschätzt werden könnte. Die Auswahl der hier zu verwendenden Kontrollen gestaltet sich als schwierig, da in der Versuchsprobe nur miRNAs vorliegen, und damit die Verwendung von Fremd-DNA als Kontrolle aufgrund eines unterschiedlichen Hybridisierungsverhaltens möglicherweise nicht geeignet ist. Zudem ist das Hinzupipettieren von Kontrollen ein fehleranfälliger Vorgang.
Yang, Dudoit et al. (2002) beschreiben eine Methode, bei der eine Mischung aller auf dem Mikroarray vorhandenen miRNA-Sequenzen in verschiedenen Konzentrationen auf den Mikroarray geprintet werden. Dieses Subset von Spots wird dann zur Normalisierung verwendet. Diese Kontrollspots haben den Vorteil, dass sie die direkte

Verteilung der miRNA-Sequenzen in den beiden Proben wiedergeben und zudem keine zusätzlichen Sequenzen in die Proben hinein gegeben werden müssten. Eine Normalisierung über dieses Subset von Kontrollen darf aber nur angewendet werden, wenn die Vorraussetzung erfüllt ist, dass nur eine geringe Anzahl von differentiell exprimierten miRNAs zwischen den Phänotypen vorliegt. Sobald ein zu großer Anteil von miRNAs differentiell exprimiert ist, würden auch die Kontrollspots mit verschoben werden und keine Null-Referenzlinie mehr bilden. Inzwischen stehen miRNA-Bibliotheken mit einer größeren Anzahl von miRNA-Sequenzen zur Verfügung z.B. von der Firma Ambion, so dass die Anzahl der Datenpunkte auf dem Mikroarray erhöht werden können und eine bessere Abschätzung der Regressionsgerade für eine Lowess-Normalisierung ermöglicht wird.

Im Rahmen der Untersuchung der N-Myc abhängigen miRNA-Expression *in vivo* und *in vitro* wurde die Mikroarray-Plattform erstmals experimentell eingesetzt. Die Ergebnisse beider Versuche zeigen mit 8 gemeinsam auftretenden miRNAs eine deutliche Überlappung auf. Zu diesen 8 gemeinsamen miRNAs gehören miR-92, mir-106a und miR17-5, welche zum miR-106a Cluster und miR-17 Cluster gehören. Für dieses Cluster wurde bereits eine c-myc abhängige differentielle Regulation in Leukämien gezeigt (O'Donnell, Wentzel et al. 2005).

Sowohl in den Vorversuchen als auch im Experiment konnte gezeigt werden, dass die miRNA Mikroarray-Plattform für die experimentelle Anwendung geeignet ist. Durch Veränderungen des Designs kann jedoch noch eine Verbesserung der Ergebnisse erzielt werden.

6 Etablierung der statistischen Auswertung eines Hochdurchsatz RNAi Screens

6.1 Einleitung

Nachdem das *Human Genom Project* zur Identifizierung nahezu aller Gene geführt hat, ist nun der nächste Schritt die Aufklärung der biologischen Funktionen der Gene. Zahlreiche genomische Ansätze sind hierzu in der Vergangenheit etabliert worden. Die meisten dieser Methoden sind indirekt und basieren auf dem Vergleich von Gensequenzen oder Expressionsmustern (Marcotte, Pellegrini et al. 1999). So agieren z.B. Gene mit dem gleichen Expressionsmuster häufig in dem gleichen biologischen Prozess. Eine Aussage über die direkte Interaktion der Gene kann hieraus jedoch nicht unmittelbar geschlossen werden. Zudem erfolgt eine Vielzahl der Regulations-mechanismen auf post-transkriptioneller Ebene, weshalb sie durch Genexpressions-studien nicht erfasst werden können.

Als effektive Methode zur Untersuchung der direkten Interaktion verschiedener Genprodukte hat sich in den letzten Jahren die RNAi Methode etabliert. (Caplen, Parrish et al. 2001; Elbashir, Harborth et al. 2001; Brummelkamp, Bernards et al. 2002; Miyagishi and Taira 2002; Paddison, Caudy et al. 2002; Paul, Good et al. 2002; Sui, Soohoo et al. 2002; Yu, DeRuiter et al. 2002).

6.1.1 Mechanismen der RNAi-Interferenz

Der RNAi Mechanismus wurde 1998 von Andrew Fire und Craig Mello erstmals publiziert (Fire, Xu et al. 1998). Sie zeigten, dass lange doppelsträngige RNA (dsRNA), die künstlich in *Caenorhabditis elegans* eingebracht wurde, zum Abbau der endogenen mRNA und somit zum Verlust des korrespondierenden Proteins führt. Hierfür erhielten sie 2006, nur 8 Jahre nach der Entdeckung, den Nobelpreis für Physiologie und Medizin (http://nobelprize.org/nobel_prizes/medicine/laureates/2006/adv.html). Dies unter-streicht die Bedeutung dieser Methode für die molekularbiologische und medizinische Forschung. Sie hat sich als Standardverfahren zum gezielten Ausschalten von Genen in Laborversuchen etabliert und ist so eine wichtige Grundlage für die systematische

Untersuchung der Funktionen der einzelnen Gene in der Zelle geworden. Es sind bisher drei verschiedene Wirkungsmechanismen bekannt.

- mRNA-Abbau
- Repression der Transkription
- Repression der Translation

mRNA Abbau

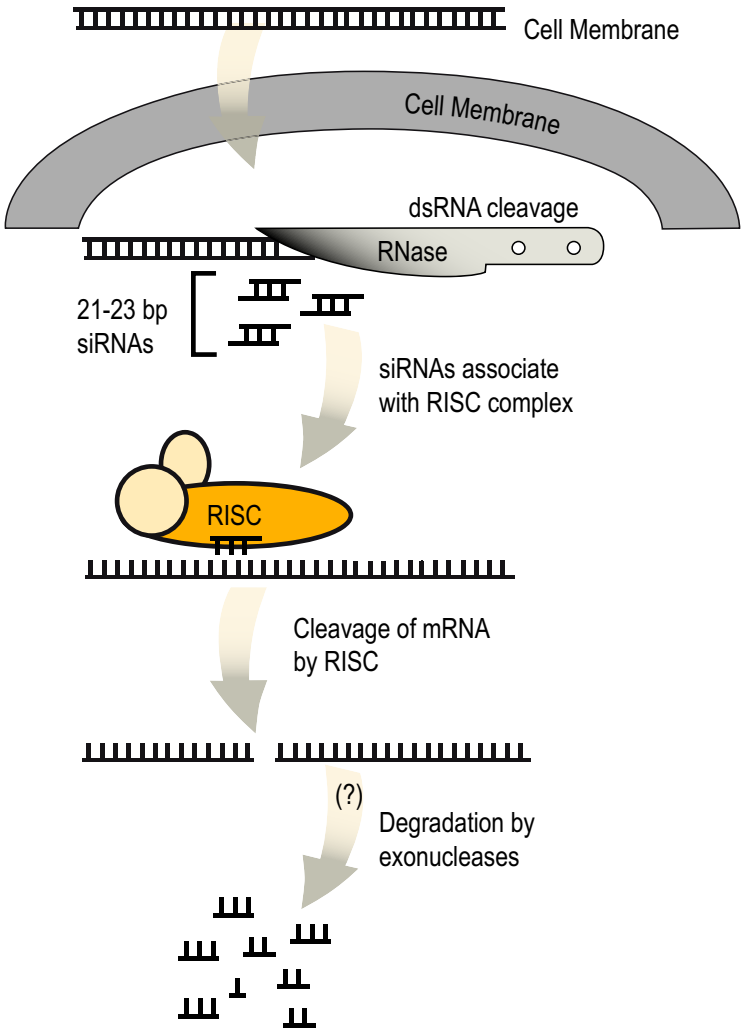


Abb. 42 siRNA induzierter mRNA Abbau. (www.ambion.com) Die mRNA wird abgebaut, indem die siRNA-Moleküle in den RISC Proteinkomplex (*RNA-induced silencing complex*) eingebaut werden und über eine komplementäre Basenpaarung an die ebenfalls im RISC integrierten RNA-Moleküle der entsprechenden mRNA binden. Die RNA-Helikase- und RNA-Nuklease-Aktivitäten des RISC-Komplexes gelangen so in die räumliche Nähe der mRNA, entwinden und spalten sie. Anschließend wird die mRNA von intrazellulären Nukleasen abgebaut

Der Abbau der mRNA erfolgt, wie in Abbildung 42 dargestellt, indem die siRNA-Moleküle in den RISC Proteinkomplex (*RNA-induced silencing complex*) eingebaut werden und über eine komplementäre Basenpaarung an die ebenfalls im RISC integrierten RNA-Moleküle der entsprechenden mRNA binden. Die RNA-Helikase- und RNA-Nuklease-Aktivitäten des RISC-Komplexes gelangen so in die räumliche Nähe der mRNA, entwinden und spalten sie. Anschließend wird die mRNA von intrazellulären Nukleasen abgebaut (Sontheimer and Carthew 2004).

Repression der Transkription

Neben dem RNA-Abbau und der Repression der Translation können siRNAs auch die Trankription von Genen reprimieren. Dies erfolgt über die Induktion einer sequenz-abhängigen lokalen Konvertierung von aktivem Euchromatin in hochkondensiertes, transkriptionell inaktives Herterochromatin. In der Spaltehefe *Schizosaccharomyces pombe* wurde ein siRNA-bindender Komplex entdeckt, der als RNA induzierte Initiation der transkriptionellen Genexpression (RIST)-Komplex bezeichnet wird (Noma, Sugiyama et al. 2004). Der Mechanismus, welcher zur Repression der Transkription führt ist in Abbildung 43 dargestellt.
Durch die Transkription repetitiver DNA-Abschnitte im Zentromeren werden partiell überlappende, nicht-kodierende RNAs generiert. Ihre Hybridisierung führt zu partiell doppelsträngigen RNA-Molekülen, die durch das Enzym DICER in siRNA-Moleküle geschnitten werden. Diese siRNA-Moleküle fungieren als Primer, die an weitere Transkripte binden und von der RNA-abhängigen RNA-Polymerase RDRP verlängert werden. Die so entstandende doppelsträngige RNA wird wieder von DICER prozessiert. Hierbei erfolgt gleichzeitig eine Amplifikation der siRNA. Diese siRNAs werden in den RITS-Komplex integriert, welcher aus drei Proteinen , Ago1, Chp1 und Tas3 besteht. Ago1 bindet die siRNA und verwendet diese als Template für die komplementäre Bindung mit RNA-Polymerase II Transkripten in der Zentrom-Region. Die Umwandlung von Eu- in Heterochromatin erfolgt durch die Methylierung des Lysins an Positon 9 von Histon H3, katalysiert durch die Histonmethyltransferase Clr4.

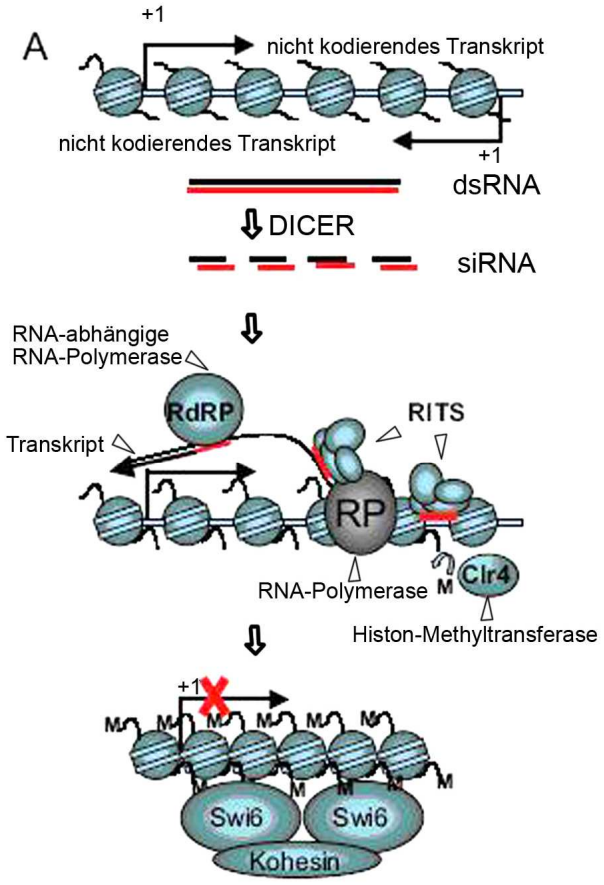


Abb. 43: Mechanismus der RNAi induzierte Repression der Transkription (Truss and Hagemeier 2004). Im Zentromer entstehen durch die Transkription repetitiver DNA-Abschnitte partiell überlappende, nicht-kodierende RNAs. Ihre Hybridisierung führt zu partiell doppelsträngigen RNA-Molekülen, die durch das Enzym DICER in in siRNA-Moleküle geschnitten werden. Die so entstandenen siRNA-Moleküle fungieren als Primer, die an Transkripte binden und von der RNA-abhängigen RNA-Polymerase RDRP verlängert werden. Die daraus resultierenden doppelsträngigen RNA wird wieder von DICER prozessiert. Hierbei erfolgt gleichzeitig eine Amplifikation der siRNA. Diese siRNAs werden dann in den RITS-Komplex integriert. In dem RITS-Komplex bindet das Protein Ago1 die siRNA und verwendet diese als Template für die komplementäre Bindung mit RNA-Polymerase II Transkripten in der Zentrom-Region. Die Umwandlung von Eu- in Heterochromatin erfolgt durch die Methylierung des Lysins an Positon 9 von Histon H3, katalysiert durch die Histonmethyltransferase Clr4.

Repression der Translation

Die Repression der Translation erfolgt über miRNAs. Der Mechanimus wird in Kapitel 5.1.1 beschrieben und ist in Abbildung 30 dargestellt.

In der Natur dient der RNAi Mechanismus zur Abwehr von Fremd-RNA z.B. von Viren (Silverstein, 1989). Der Einsatz der RNAi in höheren Organismen war lange Zeit auf Grund einer unspezifischen Interferonantwort der Zellen schwierig. 2001 wurde jedoch entdeckt, dass durch die Verwendung von kleinen dsRNA-Molekülen

(19-22 Basenpaare), den *small-interfering-RNAs* (siRNAs) oder den *short-hairpin-RNAs* (shRNAs) dieser Abwehrmechanismus unterlaufen werden kann (Elbashir, Harborth et al. 2001; Paddison, Caudy et al. 2002).

6.1.2 siRNA/shRNA-Bibliotheken

Verschiedene Forschergruppen (Bernards, Brummelkamp et al. 2006; Kimura, Wakamatsu et al. 2006; Dietzl, Chen et al. 2007) und kommerzielle Anbieter (z.B. Dharmacon, Quiagen, Proligo und Ambion) haben in den letzten Jahren RNAi-Bibliotheken generiert. Diese bestehen entweder aus chemisch synthetisierten siRNAs oder aus shRNAs, welche von viralen Vektoren exprimiert werden. Es gibt inzwischen si-/shRNAs gegen nahezu jedes Gen, womit die Möglichkeit besteht, genomweite Screens zur systematischen Untersuchung der elementaren biologischen Prozesse durchzuführen.

<u>shRNA-Bibliothek</u>

Bei den shRNA-Bibliotheken wird eine 50 Basenpaar lange Sequenz mittels eines viralen Vektors in die Zelle infiziert und dort stabil ins Genom integriert. Die siRNAs entstehen, indem das Enzym Dicer die entstehenden doppelsträngigen RNA Moleküle in kurze Fragmente zerschneidet.

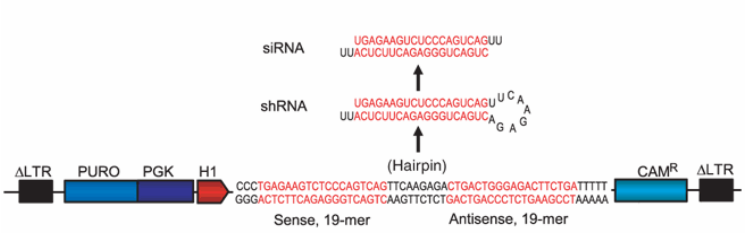


Abb. 44: Prozessierung von shRNAs (aus: Bernards, Brummelkamp et al. (2006)). Die doppelsträngigen RNA Moleküle werden durch das Enzym Dicer in siRNA-Fragmente prozessiert.

<u>siRNA-Bibliothek</u>

siRNAs bestehen aus 19 Basenpaar langen dsRNAs, die am 3´-Ende einen 2 Nukleotide langen Überhang haben. Sie entsprechen von der Struktur einer pre-miRNA und werden von der Zelle wie diese behandelt.

6.1.3 RNAi Screening

Die durch das gezielte Ausschalten einzelner Gene resultierende Veränderung des Phänotypes im subzellulären Bereich, z.B. die Veränderung der Proteinmenge oder der Translation eines Genproduktes in der Zelle, kann durch Immunfluoreszenzfärbungen mikroskopisch beobachtet werden. Durch die Verwendung verschiedener Fluoreszenz-farbstoffe können verschiedene Substrukturen parallel gemessen werden.
Die RNAi-Methode hat sich im Rahmen von Einzelversuchen als sehr effektiv erwiesen. Im Zuge der Etablierung genomweiter RNAi-Bibliotheken ist ein Bedarf an Screening-Methoden, mit welchen diese Bibliotheken vollständig getestet werden können, entstanden. So sind in den letzten Jahren Systeme entwickelt worden, die eine Umsetzung der RNAi-Methode im Hochdurchsatz-Format ermöglichen. Hierzu gehören standardisierte, zellbasierte Assays, automatisierte Mikroskopsysteme, die mit einer hohen Geschwindigkeit und Auflösung mit Fluoreszenzfarbstoffen markierte subzelluläre Strukturen aufnehmen und digitalisieren können und Programme zur Bildanalyse, die die automatisierte Extraktion und Quantifizierung verschiedener Parameter ermöglichen.

Die Umsetzung des Einzelversuches in ein Hochdurchsatz-Format, d.h. in ein 96- bzw. 384-*well* Format, stellt einige Herausforderungen an den Assay. So führt z.B. das Kultivieren von Säuger-Zellen in kleinen Volumina zu Problemen aufgrund von Nährstoffmangel, der Ansammlung von toxischen Substanzen oder Konzentrations-veränderungen infolge Verdunstung. Auch erfordert ein Screen eine wesentlich bessere Kontrolle der experimentellen Umgebungsparameter. Es gestaltet sich häufig als schwierig, die Versuchsbedingungen für alle Proben über einen längeren Zeitraum gleich gut zu gestalten bzw. zu kontrollieren. Weitere Fehlerquellen können Pipettierfehler oder unterschiedliche Konzentrationen der Reagenzien sein.

Hochdurchsatz-Screens generieren somit Datensätze, die häufig einen hohen Anteil an zufälligen und systematischen Fehlern aufweisen, was zu einer entsprechenden Anzahl von falsch positiven und falsch negativen *hits* führen kann. Wiederholungsmessungen, die zur Erhöhung der Datenqualität führen könnten, werden aufgrund der hohen Kosten und des aufwendigen Verfahrens häufig nicht oder nur in begrenzter Anzahl gemacht.

Ziel der statistischen und bioinformatischen Datenanalyse ist es, eine möglichst hohe Datenqualität zu erzielen, um so eine optimale Interpretation der Ergebnisse zu ermöglichen. Hierzu gehören, neben einer geeigneten Versuchsplanung, die Bewertung der Qualität des Screens, das Aufdecken von systematischen Fehlern, wenn möglich die Korrektur dieser Fehler, die Normalisierung der Daten und die Selektion von *hits*. Im Folgenden werden statistische Methoden zur Analyse von HTS-Daten aufgezeigt und anschließend im Rahmen der Auswertung eines RNAi-Screening Datensatzes die Etablierung einer Auswertungsroutine beschrieben und diskutiert.

6.2 Methoden zur statistische Planung und Auswertung eines Hochdurchsatz RNAi Screens

6.2.1 Experimentelles Design:

Zunächst muss ein geeignetes Versuchsdesign ausgewählt werden. Dies muss in Abhängigkeit von der biologischen Fragestellung, der verwendeten Bibliothek, aber auch in Hinblick auf die Durchführbarkeit und die statistischen Auswertungsmethoden gewählt werden.
Folgende Anforderungen werden an das Design gestellt:

Jede 96-*well* Mikrotiterplatte sollte ein Set von Kontroll-*wells* beinhalten, welche folgende Aufgaben erfüllen:

- Bewertung der Platte-zu-Platte Variabilität
- Ermittelung des Hintergrunds jeder Mikrotiterplatte anhand der Negativ-kontrollen
- Berechnung des Signalfensters, Z´-Faktors (s.u.)
- Normalisierung der Platten (s.u.)

Die Position der Kontrollen auf den Platten ist von Bedeutung, da die Randbereiche der Platten oft von externen Faktoren beeinflusst werden. Um Positionseffekte (Reihen-, Spalteneffekte) zu reduzieren, sollte die Anordnung der negativen und positiven Kontrollen so gewählt werden, dass sie im gleichen Maße in jeder Reihe bzw. Spalte verteilt sind. Dies ist jedoch oft schwierig umzusetzen, da die Bibliotheken in Mikrotiterplatten vorliegen, auf denen die ersten beiden Spalten für Kontrollen frei

gelassen sind. Die Positionierung der Kontrollen in die erste Spalte ist aber wenig geeignet, da diese *wells* die bereits erwähnten Randeffekte aufweisen und so eine hohe Fehlerrate haben. Die Umverteilung der sh-/siRNAs ist aufwendig und vor allem auch sehr fehleranfällig. Ein Kompromiss zwischen Fehleranfälligkeit und Umsetzbarkeit ist die Verlegung der Kontrollen aus der ersten Spalte in eine Spalte, die weiter in der Mitte der Platte liegt. Falls Kontrollen zur Normalisierung verwendet werden müssen, sollte eine größere Anzahl von Kontrollspots auf jeder Platte verwendet werden (s.u).
Die sh-/siRNAs sollten randomisiert verteilt werden, so dass z.B. Gene einer funktionellen Gruppe nicht alle auf einer Platte sind, sondern zufällig über die Platten verteilt werden. Hierdurch soll erreicht werden, dass jede Platte einen hohen Anteil von *Non-hits* hat, eine Voraussetzung für die Normalisierung der Daten über die Gesamtverteilung der Messwerte einer Platte (s.u.).

Wiederholungsmessungen

Zufällige Fehler sind Bestandteil von allen wissenschaftlichen Messungen. Sie treten während des gesamten Screens auf, z.B. als Folge von Spannungsveränderungen der Geräte, fehlerhaftem Pipettieren und der Reagenzien- oder Probenherstellung und –verwendung (Gunter, Brideau et al. 2003). Zufällige Fehler können über Wiederholungsmessungen kontrolliert werden. Sie erhöhen die Genauigkeit der Messungen, da die Standardabweichung mit der Zahl der Wiederholungen reduziert wird. Zudem ermöglichen sie eine direkte Abschätzung der Variabilität der Daten und reduzieren die Zahl der falsch negativen *hits*, ohne die Zahl der falsch positiven *hits* zu erhöhen. Die Variabilität von Wiederholungen innerhalb einer Platte ist geringer als die Variabilität von Wiederholungen auf verschiedenen Platten. Wiederholungsmessungen auf verschiedenen Platten werden jedoch bevorzugt, da sie eine wesentlich bessere Einschätzung der Variabilität ermöglichen, die notwendig ist, um die Ergebnisse zu generalisieren.

6.2.2 Qualitätskontrollen

Die Qualitätskontrolle ist ebenfalls ein entscheidender Bestandteil der Datenanalyse von HTS. Es gibt eine Reihe von Ursachen, die einen erheblichen Einfluss auf die Qualität der Daten eines Screens haben und zu falsch positiven bzw. falsch negativen Ergebnissen führen. Hierbei unterscheidet man zufällige und systematische Fehler.

Zufällige Fehler sind nur über Wiederholungen zu kontrollieren. Sie erzeugen Varianzen, die nur einen kleinen Einfluss auf die Verteilung der *hits* haben. Systematische Fehler können dagegen detektiert und z.T. auch mittels statistischer Verfahren korrigiert werden (Heuer C. 2002). Zu den systematischen Fehlern gehören lokale Effekte, z.B. Ecken-, Reihen- oder Spalteneffekte, die dazu führen, dass die Messwerte einzelner Reihen oder Spalten systematisch über- oder unterschätzt werden (Brideau, Gunter et al. 2003). Diese Positionseffekte treten sehr häufig in HTS auf. Sie resultieren aus Unterschieden der Temperatur, Luftfeuchtigkeit, Inkubationszeit oder Konzentrationsunterschieden der Antikörper in den verschiedenen *wells* der Platte.

Weitere Ursachen für systematische Fehler, welche auch plattenübergreifend auftreten, sind laut Heuer C. (2002) z.B.:

- Pipettierfehler (verschiedene Zelldichten, verschiedene Antikörper-Konzentrationen)
- Fehler bei der Bildauswertung

Mittels verschiedener Qualitätsparameter und diagnostischer Bilder können solche systematischen Fehler entdeckt und korrigiert werden. Zusätzlich ermöglichen die Qualitätsparameter Aussagen über die Funktionalität des Testverfahrens.

Qualitätsparameter:

Zur Bewertung von Screens werden Qualitätsparameter verwendet, die den Grad an Trennung zwischen den positiven und negativen Kontrollen eines Versuches unter Berücksichtigung der beobachteten Variabilität verwenden. Nur wenn zwischen den beiden Kontrollen ein ausreichend großes Messfenster vorliegt, sind die Voraussetzungen vorhanden, si-/shRNA Effekte zu messen.

Im Folgenden werden einige der häufig verwendeten Parameter beschrieben

- Z´-Faktor
- Signalfenster
- Signal-Hintergrund-Verhältnis
- Signal-Rausch-Verhältnis

Z´-Faktor

Der Z´-Faktor (Zhang, Chung et al. 1999) berücksichtigt neben den Mittelwerten der positiven und negativen Kontrollen die Standardabweichung beider Kontrollen.

Der Z´-faktor wird folgendermaßen berechnet:

$$Z' - \text{Faktor} = \frac{(MW_{max}\ \text{Signal} - 3\ SD_{max}) - (MW_{min}\ \text{Signal} + 3SD_{min})}{MW_{max}\ \text{Signal} - MW_{min}\ \text{Signal}}$$

MW_{max} Signal = Mittelwert der Kontrolle mit dem höheren Signal
MW_{min} Signal = Mittelwert der Kontrolle mit dem niedrigeren Signal
SD_{max} = Standardabweichung der Kontrolle mit dem höheren Signal
SD_{min} = Standardabweichung der Kontrolle mit dem niedrigeren Signal

Iversen, Eastwood et al. (2006) setzen folgende Limits für den Z´-Faktor zur Bewertung eines Versuches.

Exzellent: $Z' > 0.5$
Machbar: $0 < Z' < 0.5$
Nicht Akzeptabel: $Z' < 0$

Signalfenster (*signal window*)

Eine Alternative zum Z´-Faktor ist das Signalfenster, welches wie folgt berechnet wird:

$$\text{Signalfenster} = \frac{(MWmax\ \text{Signal} - 3\ SDmax) - (MWmin\ \text{Signal} + 3SDmin)}{SDmax}$$

MWmax Signal = Mittelwert der Kontrolle mit dem höheren Signal
MWmin Signal = Mittelwert der Kontrolle mit dem niedrigeren Signal
SDmax = Standardabweichung der Kontrolle mit dem höheren Signal
SDmin = Standardabweichung der Kontrolle mit dem niedrigeren Signal

Nach Iversen, Eastwood et al. (2006) sollte der Messwert des Signalfensters für die Durchführung eines Screens größer als 2 sein. Werte größer als 1 liegen im Toleranzbereich, bei Werten kleiner 1 sind die Unterschiede zu gering.

Signal-Hintergrund-Verhältnis

Das Verhältnis der mittleren Signalintensität zur Hintergrundintensität lässt eine Aussage über den dynamischen Bereich eines Assays zu. Er wird folgendermaßen berechnet:

$$\frac{\text{Signal}}{\text{Hintergrund}} = \frac{\text{Mittelwert Signal}}{\text{Mittelwert Hintergrund}}$$

Signal-Rausch-Verhältnis

Das Signal-Rausch-Verhältnis wird folgendermaßen berechnet:

$$\frac{\text{Signal}}{\text{Rauschen}} = \frac{\text{Mittelwert Signal} - \text{Mittelwert Hintergrund}}{\text{Standartabweichung des Hintergrundes}}$$

Das Signal-Rausch-Verhältnis lässt eine Aussage über die Glaubwürdigkeit der Signale in Bezug auf das „Hintergrundrauschen“ zu. Je größer dieses Verhältnis ist, desto leichter lassen sich Informationen extrahieren und desto zuverlässiger sind die Ergebnisse.

Diagnostische Plots

Die graphische Darstellung der Ergebnisse eines Screens ist ein wichtiges Hilfsmittel zur Bewertung der Datenqualität eines Screens. Im Folgenden werden einige der wichtigsten Darstellungen erläutert.

Boxplots der Kontrollen im Vergleich zu den si-/shRNA-Proben auf der Platte

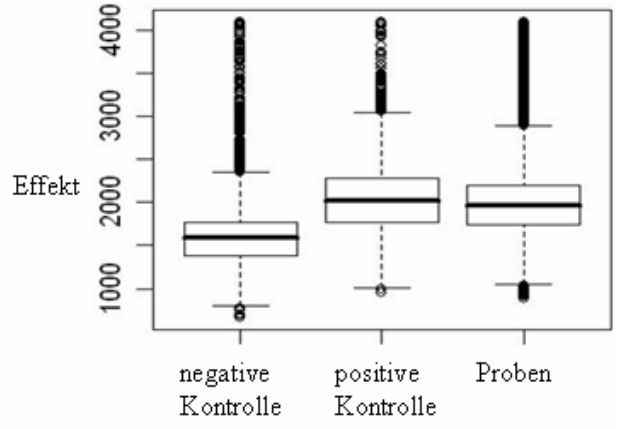


Abb. 45: Boxplots der negativen und positiven Kontrollen bzw. der si-/shRNA-behandelten Proben einer Platte. Mit den Boxplots wird die Verteilung der Messparameter von den wells in denen negativen und positive Kontrollen liegen und den wells in denen shRNA-behandelten Proben enthalten sind, dargestellt. Mittels dieser Abbildung erhält man Informationen hinsichtlich der Verteilung und Unterschiede der Kontrollen voneinander und in Relation zu den Proben. (Abbildung stammt aus Experiment 2, siehe Experimentenliste im Anhang)

In Abbildung 45 werden für jede Platte die gemessenen Parameter (z.B. die Intensität) getrennt in negative Kontrollen, positive Kontrollen und sh-/siRNAs als Boxplots aufgetragen. Hiermit erhält man einen Eindruck über die Verteilung der Datenwerte der drei Gruppen.

Boxplots der einzelnen wells einer Platte

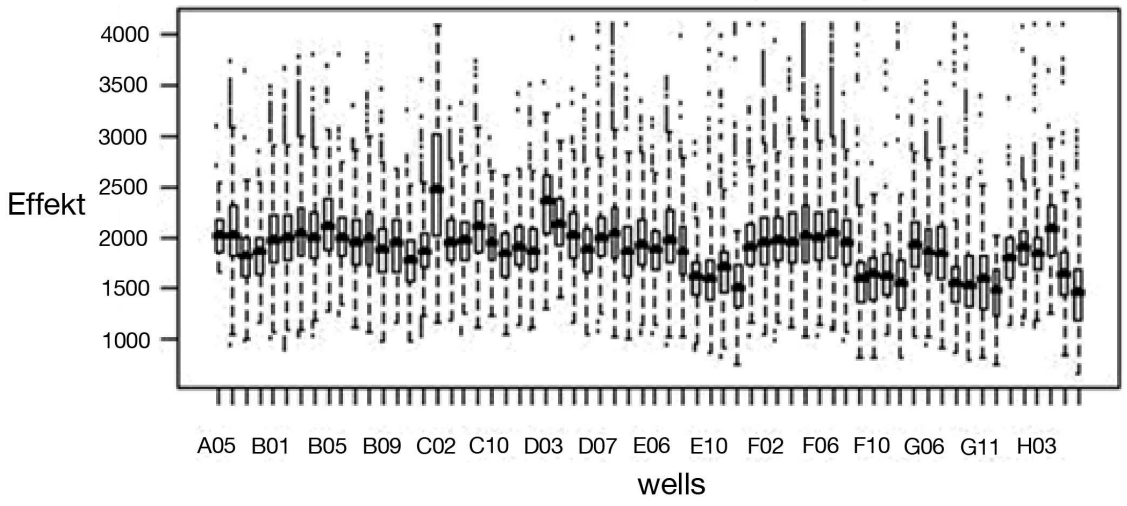


Abb. 46: Boxplot der Werte jedes *wells* einer Platte. Die Boxplots stellen die Verteilung des gemessenen Parameters der Zellen jedes *well* dar. Alle *wells* - mit Ausnahme der Kontrollen bzw. der si-/shRNAs, welche einen Effekt auslösen – sollten eine gleiche Verteilung und einen gleichen Intensitätsbereich aufweisen. Abweichungen hiervon weisen z.B. auf mögliche Spalten- bzw. Reiheneffekte bzw. Ausreißer hin. So zeigt in dieser Abbildung z.B. das *well* C2 in dieser Abbildung sowohl eine höhere Spannbreite als auch höhere gemessene Werte auf. (Abbildung stammt aus Experiment 2, siehe Experimentenliste im Anhang)

Mit der Darstellung der Werte eines jeden *wells* als Boxplot, wie sie in Abbildung 46 dargestellt ist, können z.B. „Ausreißer“ detektiert werden, bzw. erhält man einen Eindruck über die Anzahl der möglichen *hits* einer Platte, was die verwendete Normalisierungstechnik beeinflusst.

Bildplots der Platten

Bildplots stellen die gemessenen Parameter räumlich entsprechend ihrer Position auf der Platte dar. Mit Bildplots können lokale Effekte, wie Reihen-, Spalten- oder Eckeneffekte auf den Platten erkannt werden.

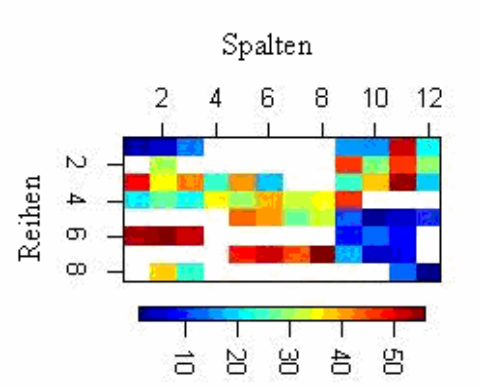


Abb. 47 Bildplot einer 96-well Mikrotiterplatte. Mit den Bildplots sollen die Platten auf lokale Effekte untersucht werden, Hierzu werden die Messwerte der einzelnen wells entsprechend ihrer Position auf der Platte farblich kodiert dargestellt. Die hierfür geltende Skala ist am unteren Bildrand dargestellt. Weiße Felder entsprechen fehlenden Messwerten (Abbildung stammt aus Experiment 3, siehe Experimentenliste im Anhang)

Darstellung von Reihen- und Spalteneffekten im Scatterplot

Zur Detektion von Reihen- und Spalteneffekten wird der Wert des gemessenen Parameters auf der y-Achse gegen die jeweilige Reihe bzw. Spalte in der x-Achse aufgetragen.

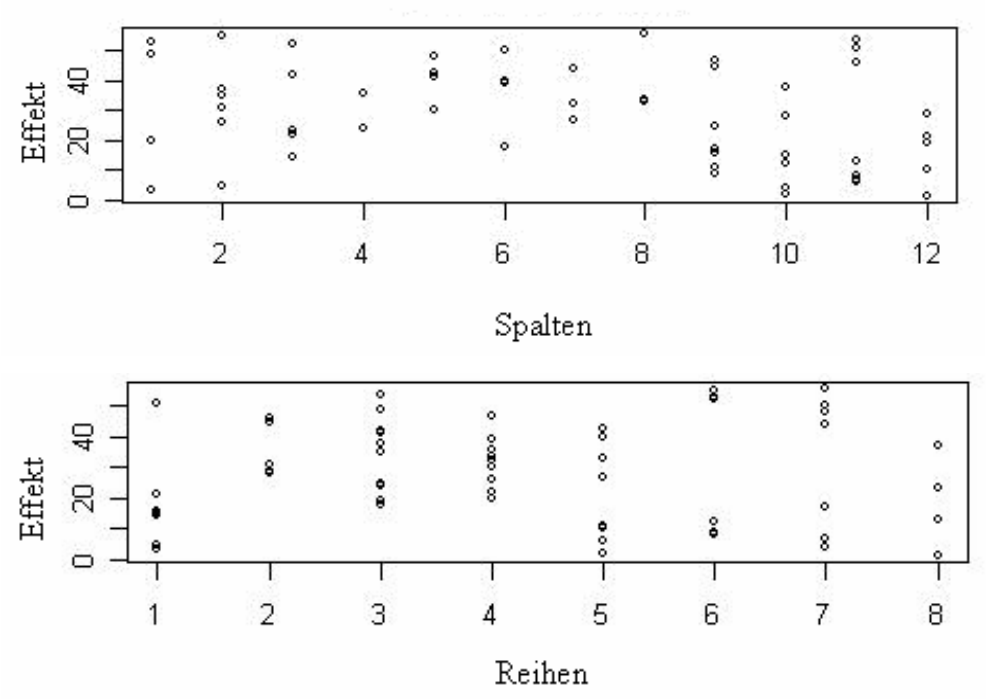


Abb. 48: Scatterplot der Spalten- bzw. Reihenummer gegen die Messwert des Parameters. Zum Auffinden möglicher Spalten- und Reiheneffekte werden die Werte jeder Platte bzw. des gesamten Screens in Abhängigkeit von der Spaltennummer (oben) und der Reihennummer (unten) dargestellt. (Abbildung stammt aus Experiment 2, siehe Experimentenliste im Anhang)

Scatterplot zum Überprüfen der Flachbildkorrektur

Je nach Position der Zelle innerhalb eines Bildes werden die gemessenen Fluoreszenzintensitäten von der Art des Lichteinfalles beeinflusst. Dieser Effekt kann durch eine so genannte Flachbildkorrektur korrigiert werden. Bei der Flachbildkorrektur wird mit Hilfe eines Objektträgers, der die Farbe des zu messenden Fluoreszenzsignals hat, ein Hintergrundbild aufgenommen. Dieses Hintergrundbild wird während der Auswertung von jedem Bild subtrahiert. Abbildung 49 zeigt die Intensitätsverteilung in Abhängigkeit von der x- bzw. y-Position ohne Flachbildkorrektur. Die Abbildung zeigt, welch großes Ausmaß dieser Effekt auf die Datenverteilung hat. Bei dieser Auswertung wurden insgesamt 3x3 Bilder aufgenommen. Ein Bogen stellt die Werte der 3 übereinander liegenden (x-Achse) bzw. der 3 nebeneinander liegenden (y-Achse) Bilder dar. Zellen, die in der Mitte der Bilder liegen, zeigen höhere Werte als Zellen im Randbereich.

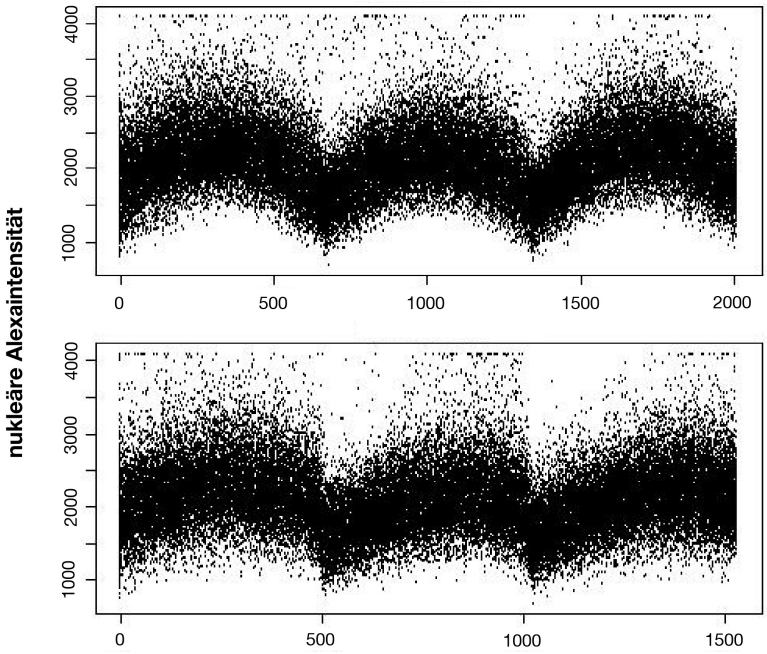


Abb. 49: Scatterplot der x- bzw. y-Position innerhalb des Bildes gegen Messwert. Bei dieser Abbildung wird für jede Zelle die Position innerhalb der *wells* – im oberen Bild in der x-Achse, im unteren Bild in der y-Achse - gegen die Intensität aufgetragen. Hiermit soll überprüft werden, ob Varianzen auftreten, die durch den Einfallswinkel des Lichtes bedingt sind. Je nach Position der Zelle innerhalb eines Bildes und damit vom Lichteinfallswinkel abhängig, erhalten die Zellen unterschiedliche Lichtmengen. Bei dieser Auswertung wurden insgesamt 3x3 Bilder aufgenommen. Ein Bogen stellt die Werte der 3 übereinander liegenden (x-Achse) bzw. der 3 nebeneinander liegenden (y-Achse) Bilder dar. Zellen, die in der Mitte der Bilder liegen, zeigen höhere Werte als Zellen im Randbereich, hieraus resultiert die Bogenstruktur. (Abbildung stammt aus Experiment 4, siehe Experimentenliste im Anhang)

Scatterplot zur Überprüfung der Reproduzierbarkeit von Wiederholungsscreens

Werden bei einem Screen biologische oder technische Wiederholungsplatten angefertigt, stellt der Grad der Übereinstimmung der Messungen ein wichtiges Kriterium für die Qualität der Platten und die Glaubwürdigkeit der Ergebnisse dar. Die Übereinstimmung kann durch Scatterplots der Werte der Wiederholungsplatten gegeneinander und der Berechnung der Korrelation zwischen den Platten erfolgen (s. Abbildung50).

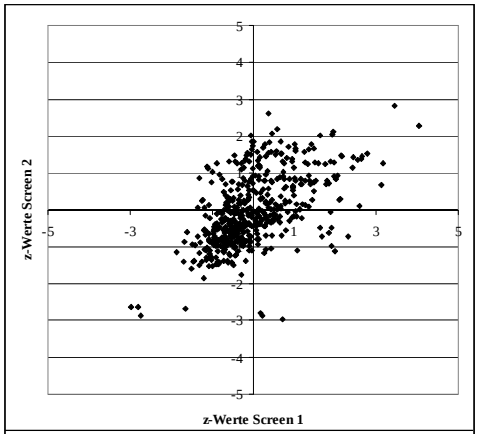


Abb. 50: Reproduzierbarkeit von zwei Screens durch Darstellung der Ergebnisse im Scatterplot. In dem Plot werden die z-Werte des Screen 1 auf der x-Achse gegen die z-Werte des Screen 2 auf der y-Achse dargestellt. Je höher die Reproduzierbarkeit ist, desto stärker liegen die Punkte entlang einer Achse, die von links unten durch den Nullpunkt nach rechts oben verläuft. (Abbildung stammt aus Experiment 2, siehe Experimentenliste im Anhang)

Normalisierung der Daten

Die Daten werden normalisiert, um systematische Fehler zu korrigieren und die Vergleichbarkeit der Messergebnisse

verschiedener Platten zu ermöglichen.

Das Normalisierungsverfahren wird anhand des Designs der Platten und der zu korrigierenden Fehler ausgewählt.

Normalisierung anhand von Kontrollspots

Die Proben einer Platte werden mittels der Kontroll-*wells* dieser Platte normalisiert. Die Normalisierung beruht auf der Annahme der zufälligen Fehlerverteilung, von der alle Werte der Platte im gleichen Maße betroffen sind. Die Position der Kontrollen ist hierbei von großer Bedeutung. Sie sollten nicht durch Reihen-, Spalten- oder

Eckeneffekte beeinflusst sein. Die Normalisierung über Kontrollen korrigiert Plattenunterschiede, jedoch keine lokalen Effekte auf den Platten.
Folgende Formel liegt der Normalisierung zugrunde (Brideau, Gunter et al. 2003)

$$x'_I = \frac{MW_{max}\ Signal - MW\ well_i}{MW_{max}\ Signal - MW_{min}\ Signal} \cdot 100\%$$

MW_{max} Signal = Mittelwert der Kontrolle mit dem höheren Signal
MW_{min} Signal = Mittelwert der Kontrolle mit dem niedrigeren Signal
MW*well*i = gemessener Wert des *wells* i
x'_I = prozentuale Erhöhung im *well* i

Die Normalisierung über Kontrollspots ist sehr fehleranfällig. So können die Kontrollen selbst sehr variabel sein, bzw. Ausreißer beinhalten. Sie sollte nur angewendet werden, wenn eine Normalisierung über die Gesamtverteilung der Daten nicht möglich ist.

Normalisierung basierend auf der Gesamtverteilung der Daten

Die Normalisierung über die Gesamtverteilung der Daten basiert auf der Annahme, dass die Mehrzahl der Proben einer Platte keine phänotypischen Effekte zeigen.
Hierbei sind verschiedene Methoden möglich:

Median-basierte Normalisierung

Die median-basierte Normalisierung (Brideau, Gunter et al. 2003) korrigiert die Daten auf den Gesamt-Median (unter vorherigem Ausschluss der Kontrollen) der Messwerte der *wells* einer Platte.

$$x'_I = 100 \cdot \frac{MW\ well_i}{Median\ aller\ wells}$$

Median aller *wells* = Median aller *wells* (ohne Kontrollen)
MW $well_i$ = gemessener Wert des $well_i$,
x'_I = prozentuale Inhibition im $well_i$.

Diese Methode sollte nur bei Hemmungs-Assays verwendet werden, bei denen die „Negativkontrolle“ hohe Werte und die *hits* niedrige Werte haben. Bei der Berechnung der normalisierten Werte steht der Gesamtmedian im Nenner. Ist er zu klein, wie es bei einem Aktivierungsassay der Fall ist, wird er zu sehr von zufälligen Faktoren z.B. unspezifischen Signalen des Hintergrundes beeinflusst, die in der Berechung zu scheinbar großen Effekten führen.

z-Wert

Die Berechnung von z-Werten basiert auf der Annahme der zufälligen Fehlerverteilung, die alle Werte der Platte im gleichen Maße betreffen. Die Probenmessungen werden hierbei re-skaliert, relativ zur Variabilität innerhalb der Platte. Die Re-Skalierung ermöglicht die Vergleichbarkeit von Datensets mit unterschiedlichem Mittelwert und Standardabweichung.

$$z_i = \frac{(MW\ well_i - MW\ aller\ wells)}{SD\ aller\ wells}$$

z_i = z-Wert des *wells*$_i$.
MW *well*$_i$ = gemessener Wert des *wells*$_i$,
MW aller *wells* = Mittelwert aller *wells* (ohne Kontrollen)
SD = Standardabweichung der Messwerte aller *wells* einer Platte (ohne Kontrollen)

Die z-Wert Transformation ermöglicht die Vergleichbarkeit verschiedener Platten, korrigiert aber nicht die lokalen Effekte.

B-Wert

Die B-Wert-Transformation (Brideau *et al*., 2003) basiert ebenfalls auf der Annahme der zufälligen Fehlerverteilung, die alle Werte der Platte im gleichen Maße betreffen. Sie entspricht weitgehend der z-Wert-Transformation, korrigiert jedoch zusätzlich Reihen- und/oder Spalteneffekte (Malo *et al*., 2006). Sie reagiert zudem resistenter gegenüber Ausreißern. Der B-Wert sollte jedoch nur dann dem z-Wert vorgezogen werden, wenn lokale Platteneffekte auftreten, da er eine geringere statistische Aussagekraft als der z-Score hat.

6.2.3 Selektion von *hits*

Es gibt verschiedene Methoden, *hits*, d.h. *wells*, deren Phänotyp sich aufgrund der si-/shRNA Behandlung verändert hat, zu selektieren.
So können die Rohdaten oder die normalisierten Daten jeder siRNA im Scatterplot aufgetragen werden. Die Selektion erfolgt anhand der Datenverteilung. Messwerte außerhalb der „Punktwolke“, deren Bereich als Datenvarianz angenommen wird, werden als *hits* selektiert.

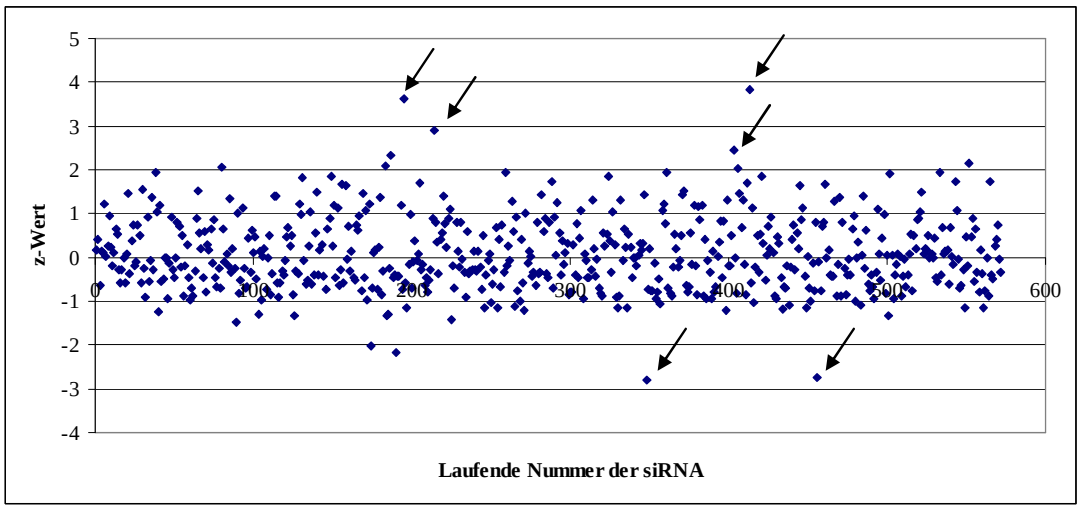


Abb. 51: Scatterplot der siRNAs auf der x-Achse gegen die z-Wert auf der y-Achse. Die siRNAs werden auf der x-Achse entsprechend ihres Auftretens im Screen, d.h. von Platte 1 fortlaufend bzw. innerhalb der Platte von *well* A1 bis H12 gegen den z-Wert auf der y-Achse aufgetragen. siRNAs deren z-Werte außerhalb der „Punktwolke" liegen werden als mögliche relevante siRNAs selektiert. Diese sind in der Abbildung mit Pfeilen markiert. (Abbildung stammt aus Experiment 3, siehe Experimentenliste im Anhang)

Das Problem dieser Methode ist, dass möglicherweise interessante siRNAs mit nur geringer oder mittlerer Effektivität nicht ausgewählt werden.

Ein weiteres Verfahren ist die Auswahl eines bestimmten Prozentsatzes an siRNAs mit der höchsten gemessenen Aktivität. Dies ist jedoch ein sehr willkürlich gewählter *cut-off*, dem ein Wahrscheinlichkeitsmodell fehlt. Auch können *hits* anhand eines auf den Werten der Kontrollen basierenden *cut-off*, z.B. die 3-fache Standardabweichung vom Mittelwert der positiven oder negativen Kontrolle, ausgewählt werden. Da der größte Teil der verwendeten siRNAs nur einen geringen bzw. keinen Effekt hat, sollte der Mittelwert und die Standardabweichung ähnlich der von den positiven Kontrollen bei Inhibitions/Antagonistenassays und ähnlich der von den negativen Kontrollen bei Aktivierungs-/Agonistenassays sein. Hier tritt jedoch ein Fehler auf, wenn Klone unterschiedliche Eigenfluoreszenz haben.

Eine weitere Möglichkeit der Selektion ist der QQ-Plot. Bei dieser Darstellung werden die Quantile einer theoretischen Verteilung gegen die empirischen Quantile von den beobachteten Merkmalswerten aufgetragen. Wenn die Merkmalswerte aus der Vergleichsverteilung stammen, stimmen die empirischen und die theoretischen Quantile überein und liegen auf einer Diagonalen.

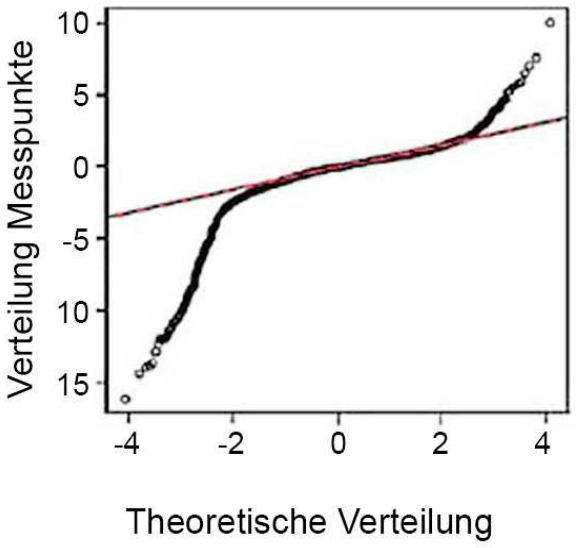


Abb. 52: QQ-Plot zur Selektion von *hits.* Der QQ-Plot vergleicht die Verteilung der Quantile der gemessenen Effekte eines Screens mit den theoretisch zu erwarteten Quantilen einer Normalverteilung, die vorliegen würde, wenn keine Effekte aufgetreten wären. siRNAs, die nicht der Normalverteilung folgen, d.h. nicht auf der in der Abbildung dargestellten roten Achse liegen, werden als mögliche relevante siRNAs selektiert (Steinbrink, T. et al. 2004)

Alle bisher beschriebenen Methoden basieren auf Screens, die nur einmal durchgeführt wurden. Wenn Wiederholungen von Screens vorliegen, kann die Reproduzierbarkeit bzw. Variabilität der Wiederholungsmessungen mit in Betracht gezogen werden. Ähnlich der Auswertung von Mikroarray-Experimenten liegen hier Daten mit einer Vielzahl von gemessenen Parametern (si-/shRNAs) bei einer nur geringen Anzahl von Wiederholungsmessungen vor. Eine Methode, die für die Auswertung von Mikroarrays mit einer nur geringen Zahl von Wiederholungen etabliert worden ist, ist der *Volcanoplot* (s. Kapitel 2.2.7), bei welchem der gemittelte Messwert einer si-/shRNA gegen die absolute t-Statistik dieser si-/shRNA aufgetragen wird. Selektiert werden si-/sh-RNAs, die sowohl bei dem gemittelten Messwert als auch bei der absoluten t-Statistik über Schwellenwerten liegen, die vom Anwender anhand der Datenverteilung ausgewählt werden.

6.3 Auswertung eines shRNA-Screening Datensatzes

6.3.1 Datensatz

In dem Screen, der im Rahmen dieser Arbeit zur Etablierung der Datenanalysemethoden verwendet worden ist, sollte der Einfluss verschiedener Kinasen auf die Stabilität von c-Myc getestet werden. Hierzu wurde eine shRNA-Kinase-Bibliothek des *Netherlands*

Cancer Institute (NKI, Amsterdam) (Bernards, Brummelkamp et al. 2006) verwendet. Diese besteht aus 638 shRNA Pools. Jeder Pool beinhaltet 3 verschiedene shRNA Sequenzen gegen die gleiche Kinase. Es werden jeweils 3 shRNAs verwendet, um die unterschiedliche Effektivität der einzelnen shRNAs auszugleichen. Als Vektor für die shRNAs fungiert pRetroSuper, der eine Puromycin-Resistenz hat.

6.3.2 Experimentelle Durchführung des Screens

Die Experimentelle Durchführung des Screens ist in Abbildung 53 dargestellt. Phönixzellen werden in einer 96-*well* Platte herangezogen und mit den pRetroSupervektoren der Bibliothek infiziert. Die von den Phönixzellen produzierten Retroviren werden randomisiert auf 8 x 96-*well*-Platten verteilt, in denen sich bereits U2OS-Zellen befinden. Zusätzlich enthält jede Platte 5 positive Kontrollen und 5 negative Kontrollen. Als positive Kontrollen werden mit *shScrambled* behandelte U2OS-Zellen verwendet, als negative Kontrollen U2OS-Zellen, bei denen der 1. Antikörper nicht zugegeben wurde. Der gesamte Screen wurde 3 mal wiederholt. Screen1 und Screen 3 haben die gleiche Verteilung der shRNAs auf die 8 x 96-*well*-Platten, bei Screen 2 ist eine andere Verteilung vorgenommen worden. Eine Selektion der Zellen, die erfolgreich infiziert wurden, erfolgt über die im Vektor integrierte Puromycin-Resistenz. Nach dem Fixieren werden die Zellen mit dem Hoechst-Farbstoff gefärbt. Dieser bindet an die DNA im Zellkern und wird bei der anschließenden Bildanalyse zum Auffinden der Zellkerne verwendet. Anschließend werden die Zellen zunächst mit einem Primärantikörper gegen c-Myc und dann mit einem Sekundärantikörper an den der Fluoreszenzfarbstoff Alexa-488 gekoppelt ist, gefärbt. Der Versuch wurde von Theresia Kress am IMT in Marburg durchgeführt.

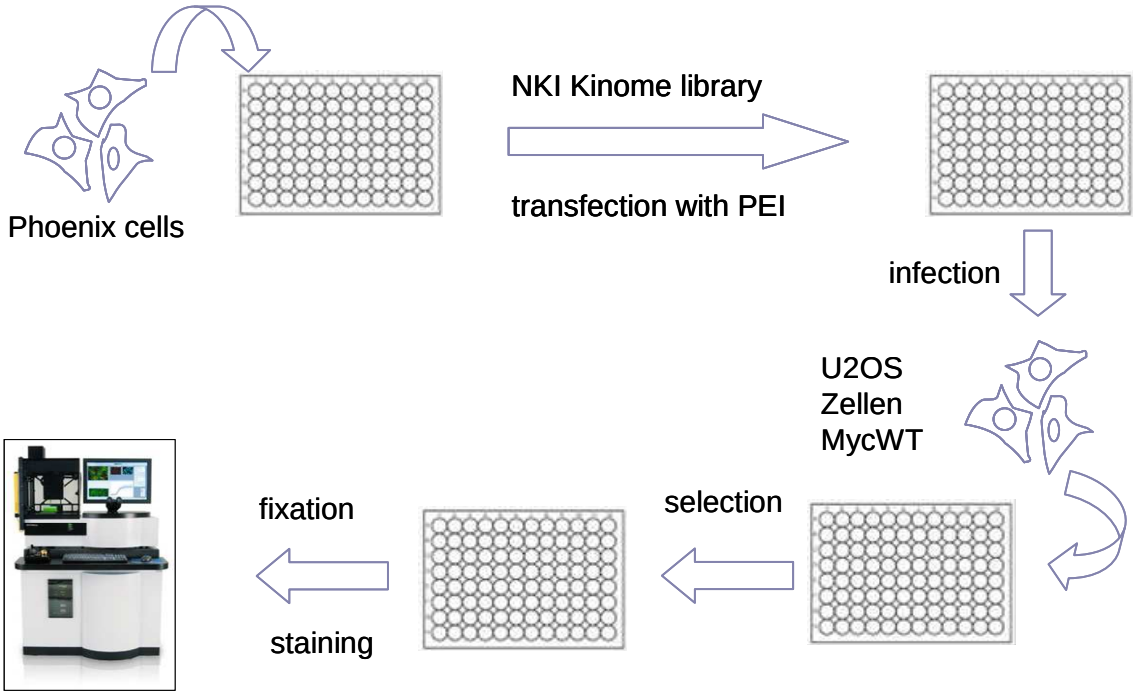


Abb. 53: Schematische Darstellung der experimentellen Durchführung des shRNA Screens. Phönixzellen werden in einer 96-*well* Platte herangezogen und mit den pRetroSupervektoren der NKI Kinase Bibliothek infiziert. Die von den Phönixzellen produzierten Retroviren werden randomisiert auf 8 x 96-*well*-Platten verteilt, in denen sich bereits U2OS-Zellen befinden. Eine Selektion der Zellen, die erfolgreich infiziert wurden, erfolgt über eine Puromycin-Resistenz. Die Zellen werden fixiert und der Zellkern mit Höchst-Farbstoff angefärbt. Anschließend werden die Zellen zunächst mit einem Primärantikörper gegen c-Myc und dann mit einem Sekundärantikörper an den der Fluoreszenzfarbstoff Alexa-488 gekoppelt ist, gefärbt. Neben den *wells* mit shRNA behandelten Zellen enthält jede Platte 5 positive Kontrollen und 5 negative Kontrollen. Als positive Kontrollen werden mit *shScrambled* behandelte U2OS-Zellen verwendet, als negative Kontrollen U2OS-Zellen, bei denen der 1. Antikörper nicht zugegeben wurde.

Die Aufnahme der Platten erfolgte mit dem BD Pathway Imager$_{TM}$ (s. Abbildung 54). Hierbei handelt es sich um ein Fluoreszenzmikroskop, das automatisiert Mikroskopbilder der einzelnen *wells* einer gesamten Platte, nach vorheriger Anregung der Fluoreszenzfarbstoffe mit Licht der entsprechenden Wellenlänge, aufnimmt.

Abb. 54: BD Pathway Imager$_{TM}$

Für beide Farbstoffe wurden von jedem *well* 9 Einzelbilder aufgenommen.

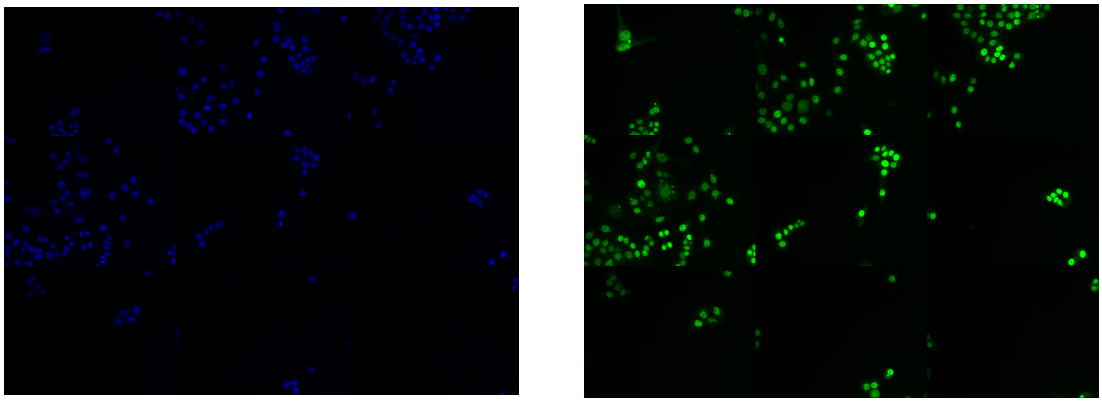


Abb. 55: Mikroskopbilder eines *wells.* Auf der linken Seite sind die mit dem Höchstfarbstoff angefärbten Zellkerne zu sehen, auf der rechten Seite der Farbstoff Alexa, mit welchem die c-myc-Level in den Zellkernen gemessen wird.

Mit der Software Attovision werden diese Bilder analysiert und für die Auswertung relevante Parameter berechnet. Die Software identifiziert in einem ersten Schritt in jedem *well* die einzelnen Zellkerne anhand der Höchstfärbung der DNA. Anschließend wird über die Anregung des Farbstoffes Alexa die Intensität und die Körnigkeit/Textur (Granularität) dieses Farbstoffes im Zellkern gemessen. Diese Parameter wurden ausgewählt, da sie zu diesem Zeitpunkt beide als mögliches Maß für das c-Myc-Level in der Zelle in Betracht gezogen wurden. Die hierbei generierten Daten dienten als Basis für die weitere statistische Auswertung.

6.3.3 Vorversuch zur Etablierung eines Assays zur Messung der c-Myc Stabilität nach Kinase-*knockdown*

Im Vorfeld jedes Screens müssen zunächst die experimentellen Versuchsbedingungen ausgetestet werden. Dies erfolgt zunächst in Einzelversuchen. Mit einem Vorversuch im Hochdurchsatz-Format soll dann getestet werden, ob sich die Bedingungen für die Durchführung des Screens eignen.

Design eines Vorversuches

Für den Vorversuch wurden 3 Testplatten nach dem folgenden Schema erstellt:

Platte 1

	1	2	3	4	5	6	7	8	9	10	11	12
A	max	mid	low	max	mid	low	max	mid	low	max	mid	low
B	max	mid	low	max	mid	low	max	mid	low	max	mid	low
C	max	mid	low	max	mid	low	max	mid	low	max	mid	low
D	max	mid	low	max	mid	low	max	mid	low	max	mid	low
E	max	mid	low	max	mid	low	max	mid	low	max	mid	low
F	max	mid	low	max	mid	low	max	mid	low	max	mid	low
G	max	mid	low	max	mid	low	max	mid	low	max	mid	low
H	max	mid	low	max	mid	low	max	mid	low	max	mid	low

Platte 2

	1	2	3	4	5	6	7	8	9	10	11	12
A	low	max	mid	low	max	mid	low	max	mid	low	max	mid
B	low	max	mid	low	max	mid	low	max	mid	low	max	mid
C	low	max	mid	low	max	mid	low	max	mid	low	max	mid
D	low	max	mid	low	max	mid	low	max	mid	low	max	mid
E	low	max	mid	low	max	mid	low	max	mid	low	max	mid
F	low	max	mid	low	max	mid	low	max	mid	low	max	mid
G	low	max	mid	low	max	mid	low	max	mid	low	max	mid
H	low	max	mid	low	max	mid	low	max	mid	low	max	mid

Platte 3

	1	2	3	4	5	6	7	8	9	10	11	12
A	mid	low	max	mid	low	max	mid	low	max	mid	low	max
B	mid	low	max	mid	low	max	mid	low	max	mid	low	max
C	mid	low	max	mid	low	max	mid	low	max	mid	low	max
D	mid	low	max	mid	low	max	mid	low	max	mid	low	max
E	mid	low	max	mid	low	max	mid	low	max	mid	low	max
F	mid	low	max	mid	low	max	mid	low	max	mid	low	max
G	mid	low	max	mid	low	max	mid	low	max	mid	low	max
H	mid	low	max	mid	low	max	mid	low	max	mid	low	max

Abb. 56: Design eines Vorversuches zum Austesten der Stabilität der Versuchsbedingungen im Hochdurchsatz-Format
max – Probe, die einen maximalen Effekt erwarten lässt
mid - Probe, die einen mittleren Effekt erwarten lässt
low - Probe, die einen minimalen Effekt erwarten lässt

Als shRNAs wurden *shScrambled* verwendet, um *wells* mit maximalen Werten, d.h unveränderten c-Myc Level zu erhalten und eine shRNA gegen c-Myc um *wells* mit mittleren Werten zu erhalten, da keine vollständige Abschaltung von c-Myc zu erwarten ist. In den *wells*, in denen minimale Werte erzielt werden sollten, wurde kein Primärantikörper zugefügt. Die drei Testplatten wurden jeweils an zwei aufeinander folgenden Tagen gemessen.

Qualitätsparameter und diagnostische Plots

Folgende Parameter wurden im Rahmen des Vorversuches getestet:

Messfenster

Auf den Testplatten befindet sich eine große Anzahl von negativen und positiven Kontrollen. Hieraus kann der Z´-Faktor bzw. das Signalfenster berechnet werden, was einen Anhaltspunkt über die Unterschiede gibt, welche mit den gewählten Versuchsbedingungen gemessen werden können. Die zusätzliche Verwendung einer Probe (shRNA gegen c-Myc), die eine „mittlere“ Intensität zwischen positiver und negativer Kontrolle aufzeigt, zeigt die Robustheit der Messungen in einem Messbereich, der den möglichen „*hits*“ entspricht.

Räumliche Effekte

Die Anordnung der verschiedenen Proben auf den Testplatten lässt Schlüsse auf mögliche Reihen-, Spalten- und Eckeneffekte zu.

Zeitliche Effekte

Die Messung der Platten an zwei aufeinander folgenden Tagen lässt Schlüsse auf die Stabilität des Experimentes zu. Dies ist wichtig, da die Durchführung eines Screens oft mehrere Tage dauert.

Qualitätsparameter

Für jede Platte werden die nukleäre Granularität und Intensität des Farbstoffes Alexa betrachtet. Zunächst werden Qualitätsparameter berechnet (s. Tabelle 15), die Aussagen über den Messbereich und das Signal-Rauschen-Verhältnis bzw. Signal-Hintergrund-Verhältnis zulassen.

Die Z´-Faktoren liegen bei der Messung der Intensität etwas geringer als bei der Granularität. Beide Parameter zeigen Werte über Null und zum Teil über 0,5. Die Z´-Faktoren der Messungen am zweiten Tag sind niedriger als die jeweilige Platte am ersten Tag. Das Verhältnis Signal zu Hintergrund liegt bei beiden Parametern um 2. Das Signal-Rausch-Verhältnis ist sowohl bei den minimalen Werten, als auch bei den maximalen Werten sehr hoch.

Tab. 15: Qualitätsparameter für die Granularität (a) und die Intensität (b) des Alexafarbstoffes im Zellkern

a) Granularität

	Platte_1_1	Platte 1_2	Platte 2_1	Platte 2_2	Platte 3_1	Platte 3_2
Median der positiven Kontrollen	191.10	176.72	186.09	171.37	181.02	161.94
Standardabweichung der pos. Kontrollen	10.39	13.16	15.99	17.97	16.83	17.32
Median der negativen Kontrollen	80.22	80.42	81.85	80.89	80.44	79.53
Standardabweichung der neg. Kontrollen	0.74	1.64	0.89	1.46	0.94	1.14
Z´-Faktor (median)	0.70	0.54	0.51	0.36	0.47	0.33
Signalfenster (median)	8.19	4.22	3.65	1.94	2.98	1.64
maximale Signal-Rausch-Verhältnis	18.39	13.42	11.63	9.54	10.76	9.35
minimale Signal-Rausch-Verhältnis	107.91	49.12	91.65	55.34	85.33	69.75
Signal-Hintergrund-Verhältnis	2.38	2.20	2.27	2.12	2.25	2.04

b) Intensität

	Platte_1_1	Platte 1_2	Platte 2_1	Platte 2_2	Platte 3_1	Platte 3_2
Median der positiven Kontrollen	1144.84	1183.04	1004.55	1006.37	1123.46	1107.88
Standardabweichung der pos. Kontrollen	63.19	69.99	57.11	55.68	99.11	95.38
Median der negativen Kontrollen	583.06	636.48	501.11	543.60	566.72	603.98
Standardabweichung der neg. Kontrollen	12.47	24.64	16.61	16.36	19.96	20.54
Z´-Faktor (median)	0.60	0.48	0.56	0.53	0.36	0.31
Signalfenster (median)	3.98	2.84	3.93	3.88	1.88	1.58
maximale Signal-Rausch-Verhältnis	18.12	16.90	17.59	18.07	11.34	11.62
minimale Signal-Rausch-Verhältnis	46.76	25.83	30.18	33.23	28.39	29.40
Signal-Hintergrund-Verhältnis	1.96	1.86	2.00	1.85	1.98	1.83

Die Werte für den Z´-Faktor und das Signalfenster zeigen, dass das Messfenster für die Durchführung eines Screens ausreichend groß ist. Die Werte des Signal-zu-Hintergrund-Verhältnisses zeigen, dass die Signale deutlich über dem Hintergrund liegen und die Signal-zu-Rauschen-Werte deuten auf eine gute „Verlässlichkeit“ der Signalintensitäten.

Diagnostische Plots

Um zu sehen, ob die Intensität von c-Myc sich linear zur Granularität verhält, werden die Wertepaare für jede shRNA im Scatterplot dargestellt (s. Abbildung 57).

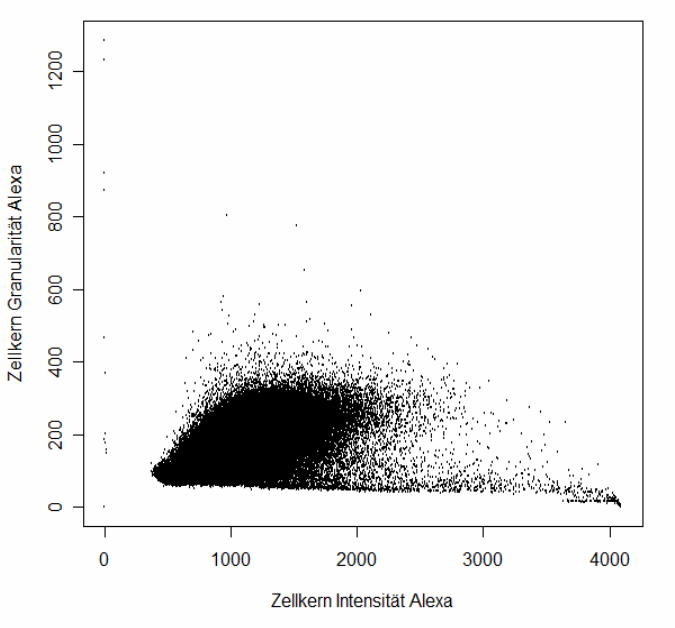


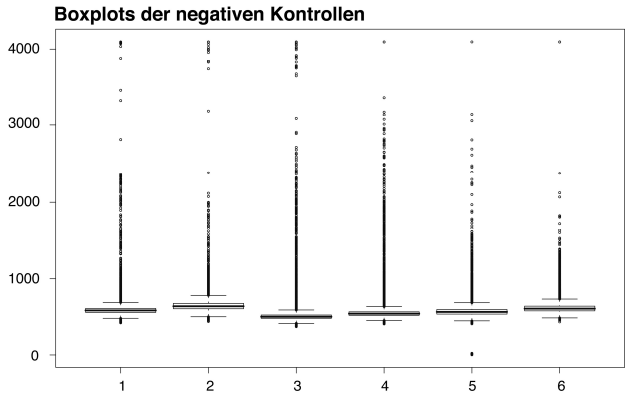
Abb. 57: Scatterplot der Intensität gegen die Granularität des nukleären Alexafarbstoffes. In dem Scatterplot wird die Intensität des nukleären Alexafarbstoffes auf der x-Achse gegen die Granularität auf der y-Achse aufgetragen. Bei niedrigen Intensitäten steigt die Granularität mit zunehmender Intensität, bei einer Intensität von ~1000 tritt jedoch eine Sättigung der Granularität ein.

Die Abbildung zeigt, dass mit zunehmender Intensität die Messwerte für die Granularität nicht weiter steigen, d.h., dass die steigenden c-Myc-Level anhand der Textur der Oberfläche ab einem bestimmten Level nicht mehr linear dargestellt werden können und sich somit als Messparameter nicht eignen.

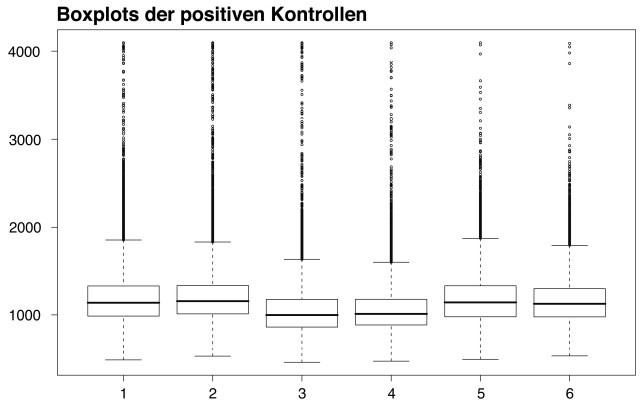
Boxplots der Kontrollen

In Abbildung 58 ist die Verteilung der drei verschiedenen Kontrollen auf den verschiedenen Platten des Vorversuches darstellt. Hier wird noch einmal bildlich dargestellt, dass deutliche Unterschiede zwischen negativen und positiven Kontrollen vorliegen. Die „mittleren“ Kontrollen befinden im Bereich zwischen den positiven und negativen Kontrollen.

(a)



(b)



(c)

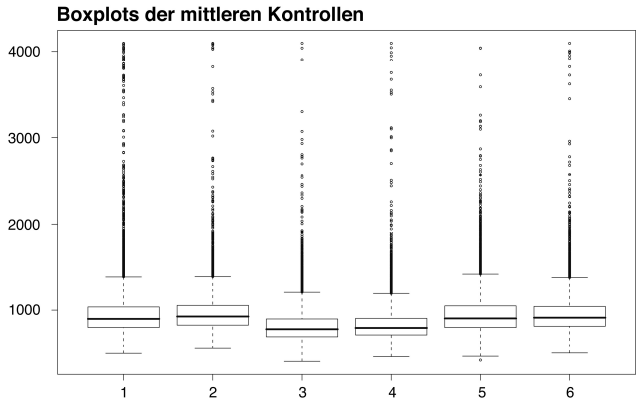


Abb. 58: Boxplots der negativen, positiven und „mittleren“ Kontrollen der Platten des Kontrollversuches. Die Boxplots geben die Verteilung der mittleren nukleären Alexaintensität der *wells* der negativen Kontrollen (a), der positiven Kontrollen (b) und der „mittleren“ Kontrollen (c) jeder Platte an. Die Werte der drei Kontrollen liegen bei allen drei Platten in gleichen Intensitätsbereichen und weisen eine ähnliche Spannbreite auf. Die positiven Kontrollen liegen deutlich höher als die negativen Kontrollen, die „mittleren“ Kontrollen zwischen den positiven und negativen Kontrollen.

Bildplots

Von jeder Platte wurden Bildplots für die Anzahl der gemessenen Zellen pro *well* und der medianen nukleären Alexa-Intensität erstellt. Diese Plots sollen mögliche räumliche Effekte auf den Platten darstellen.

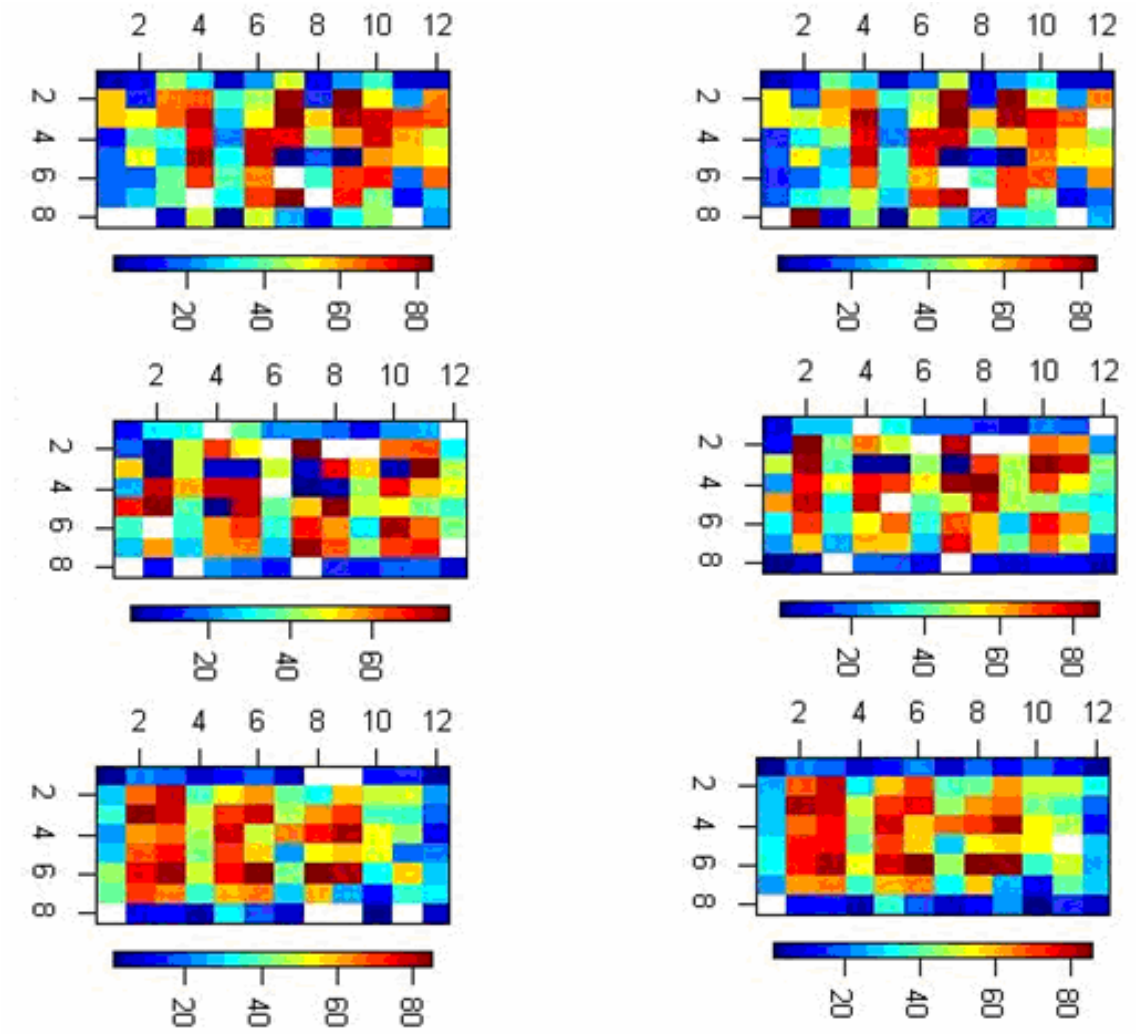


Abb. 59: Bildplots der Zellzahlen der 6 Platten des Vorversuches. Mit den Bildplots soll geprüft werden ob die Anzahl der Zellen in den einzelnen *wells* positionsabhängig ist. Hierzu wird die Anzahl der Zellen in den einzelnen *wells* entsprechend ihrer Position auf der Platte farblich kodiert dargestellt. Weiße Felder sind *wells*, die aufgrund einer niedrigen Zellzahl (< 50 Zellen) von der Auswertung ausgeschlossen wurden. Die drei Platten, die am ersten Tag gemessen wurden, sind untereinander in der linken Spalte dargestellt, die jeweilige dazugehörende Messung vom zweiten Tag in der gleichen Reihe in der rechten Spalte. Bei allen Platten ist in den Randbereichen die Anzahl der Zellen niedriger als in der Mitte der Platte.

Die Abbildung 59 zeigt, dass die Anzahl der gemessenen Zellen an den Rändern niedriger ist als in der Mitte der Platte. Diese räumlichen Effekte verstärken sich bei der Messung am zweiten Tag nicht.

Mediane Intensitäten des nukleären Alexafarbstoffes

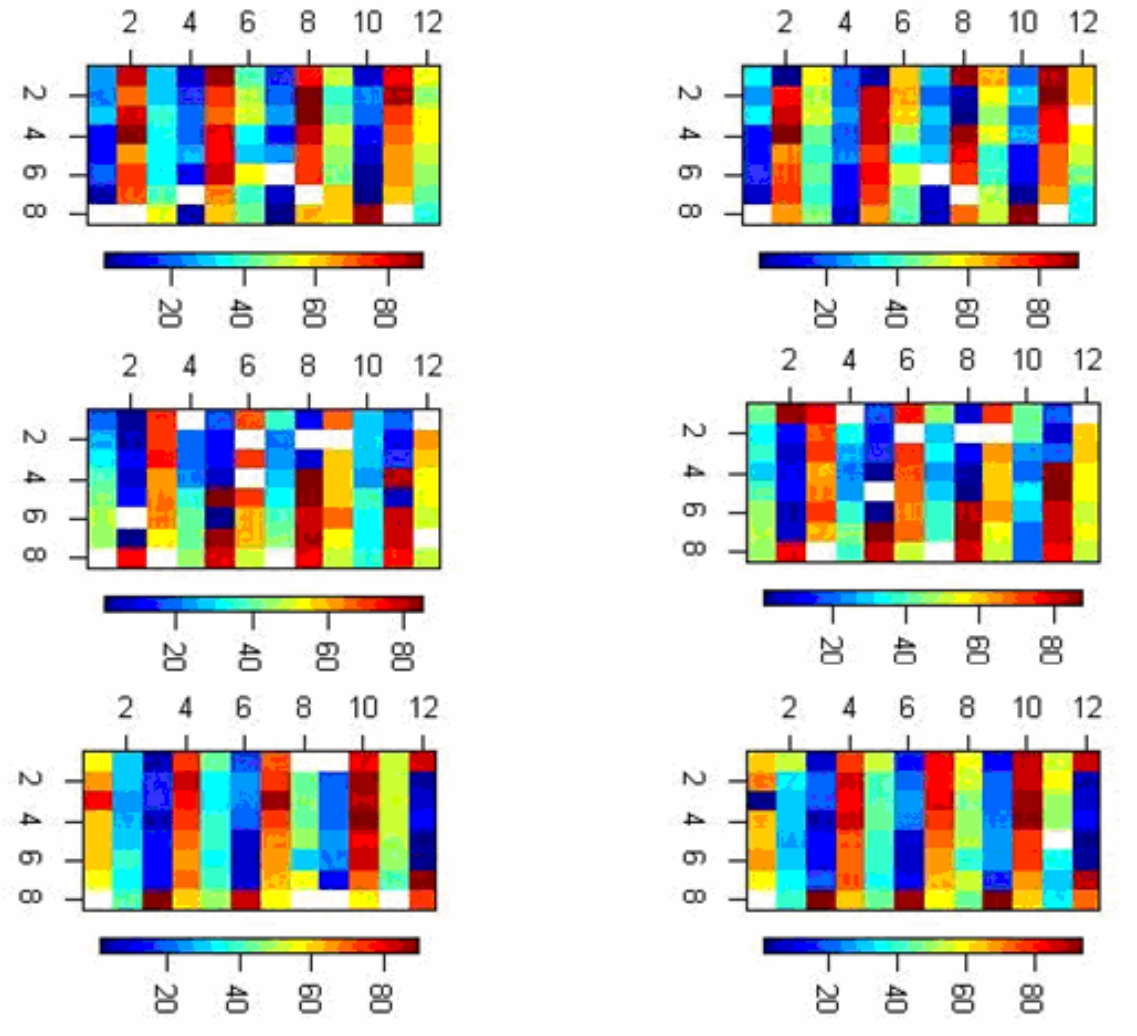


Abb. 60: Bildplots der Intensität des nukleären Alexa-Farbstoffes. Mit den Bildplots soll geprüft werden, ob positionsabhängige Effekte des nukleären Alexa-Farbstoffes auftreten. Hierzu wird die mittlere Intensität des nukleären Alexafarbstoffes für die einzelnen *wells* entsprechend ihrer Position auf der Platte farblich kodiert dargestellt. Weiße Felder sind *wells*, die aufgrund einer niedrigen Zellzahl (< 50 Zellen) von der Auswertung ausgeschlossen wurden. Die drei Platten, die am ersten Tag gemessen wurden, sind in der linken Spalte dargestellt, die jeweilige dazugehörende Messung vom zweiten Tag in der gleichen Reihe in der rechten Spalte. Die Verteilung der drei Kontrollen, welche spaltenweise angeordnet sind, ist gut zu sehen. Innerhalb der Spalten treten Varianzen auf, bei einem Teil der Spalten nehmen die Intensitäten von oben nach unten ab oder auch zu. Es treten kaum Unterschiede zwischen den Messungen am ersten und am zweiten Tag auf.

Die Intensitäten des Alexa-Farbstoffes im Zellkern entsprechen den zu erwartenden c-Myc Leveln. Da die 3 Kontrollen – mit den entsprechend unterschiedlichen Intensitäten spaltenweise angeordnet sind, zeichnet sich ein entsprechendes Muster ab. Innerhalb der Spalten sind jedoch auch Effekte zu beobachten, d.h. in einigen der Spalten erfolgt eine systematische Veränderung der Intensitäten von oben nach unten, die in beide Richtungen, d.h. als Zunahme der Intensitäten bzw. als Abnahme der Intensitäten

erfolgen kann. Es treten kaum Unterschiede zwischen den Messungen am ersten und am zweiten Tag auf.

Detektion von Reihen- und Spalten-Effekten

Um mögliche Reihen- und Spalten-Effekte noch besser erkennen zu können, wurden für jede Platte die Intensitätswerte jeder Reihe bzw. Spalte im Scatterplot dargestellt.

Reiheneffekte

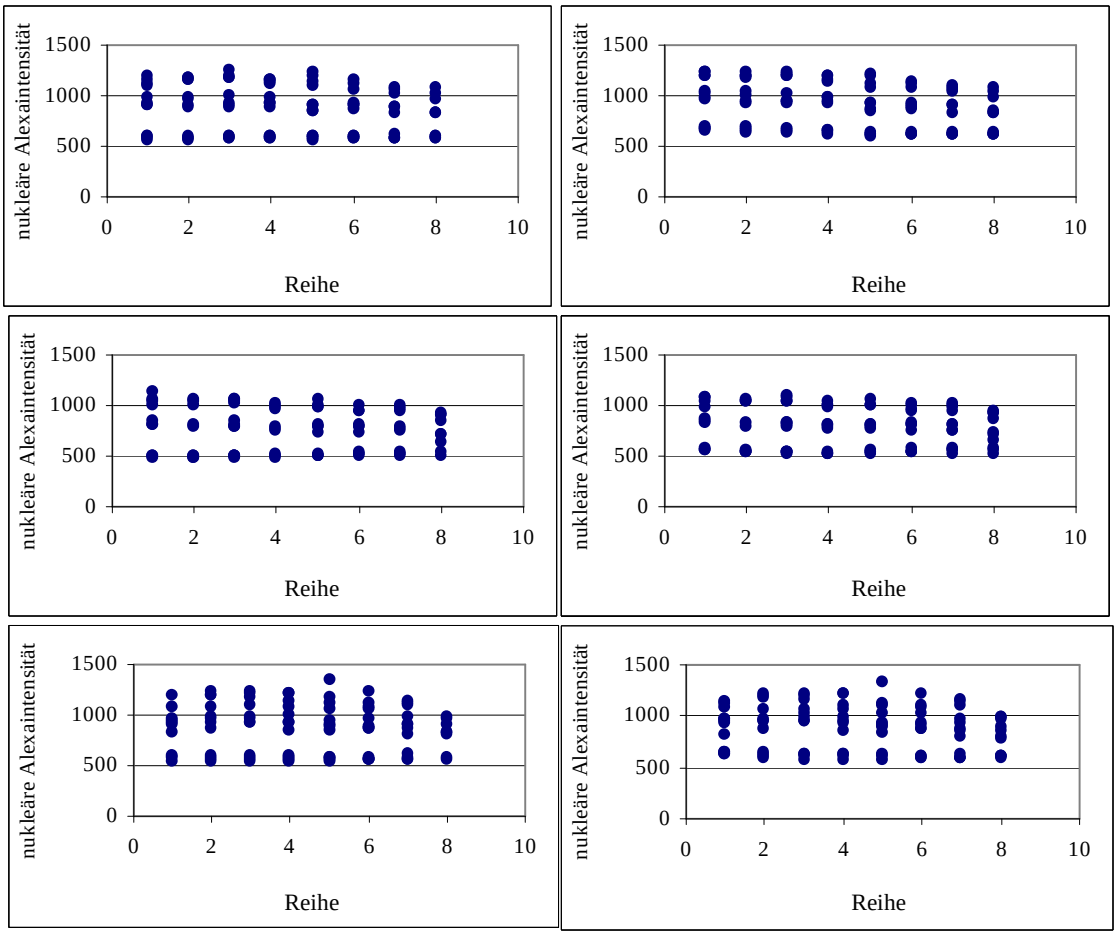


Abb. 61: Scatterplot Reihe gegen nukleäre Alexaintensität der Testplatten. In dieser Abbildung wird für jede Platte die Reihennummer (x-Achse) gegen die nukleäre Alexaintensität aufgetragen. Mit der Darstellung sollen mögliche Reiheneffekte aufgezeigt werden. Diese sind z.B. auf Platte 2 und 3 in Reihe 8 zu sehen, da die Werte im Vergleich zu den anderen Reihen höher liegen.

Da alle drei Proben in jeder Reihe gleichmäßig vertreten sind, erwarten wir eine gleichmäßige Verteilung der Werte. Es fällt bei fast allen Platten auf, dass Reihe 8 insgesamt höhere Werte hat als die anderen Reihen. Die Werte der Platten 2_1 bzw. 2_2 steigen mit zunehmender Reihennummer an.

Spalteneffekte

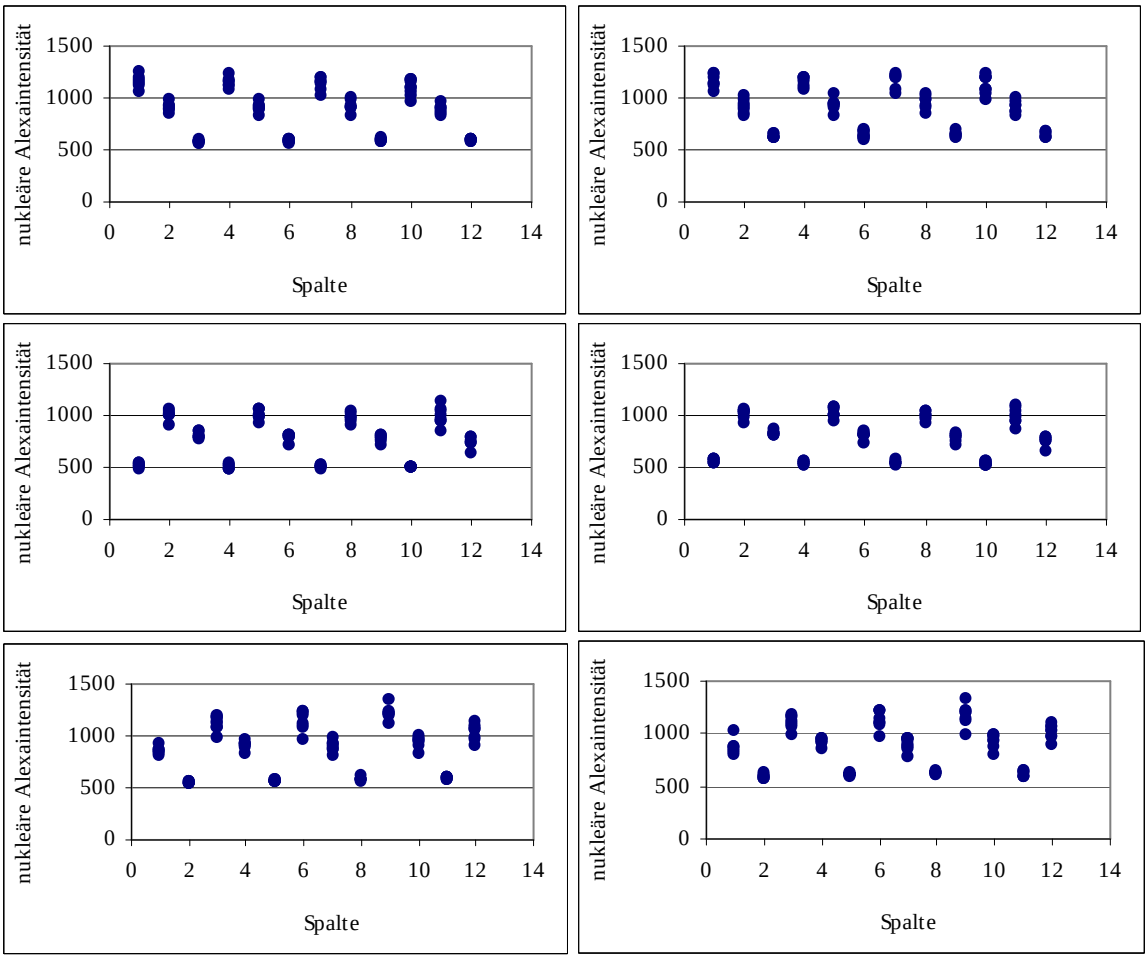


Abb. 62: Scatterplot Spalte gegen Intensität des Alexafarbstoffes im Zellkern der Testplatten. In dieser Abbildung wird für jede Platte die Spaltennummer (x-Achse) gegen die nukleäre Alexaintensität (y-Achse) aufgetragen. Da die verschiedenen Kontrollen spaltenweise angeordnet sind, findet sich auch hier die Verteilung der Intensitäten wieder.

Da pro Spalte eine Probe aufgetragen ist, erwarten wir bei der spaltenabhängigen Darstellung der Intensitäten entsprechende „Sprünge“. Diese sind deutlich zu beobachten.

Überprüfung der Flachbild-Korrektur

Vor der Bildanalyse wurde eine Flachbild-Korrektur vorgenommen. Die folgenden Abbildungen, in welchen die Positionen der Zellen auf der x-Ebene bzw. auf der y-Ebene innerhalb der aufgenommenen (3x3) Bilder gegen die Intensität aufgetragen sind, zeigen, dass kaum noch *well*-Positions abhängige Effekte auftreten, somit eine ausreichende Korrektur erfolgt ist.

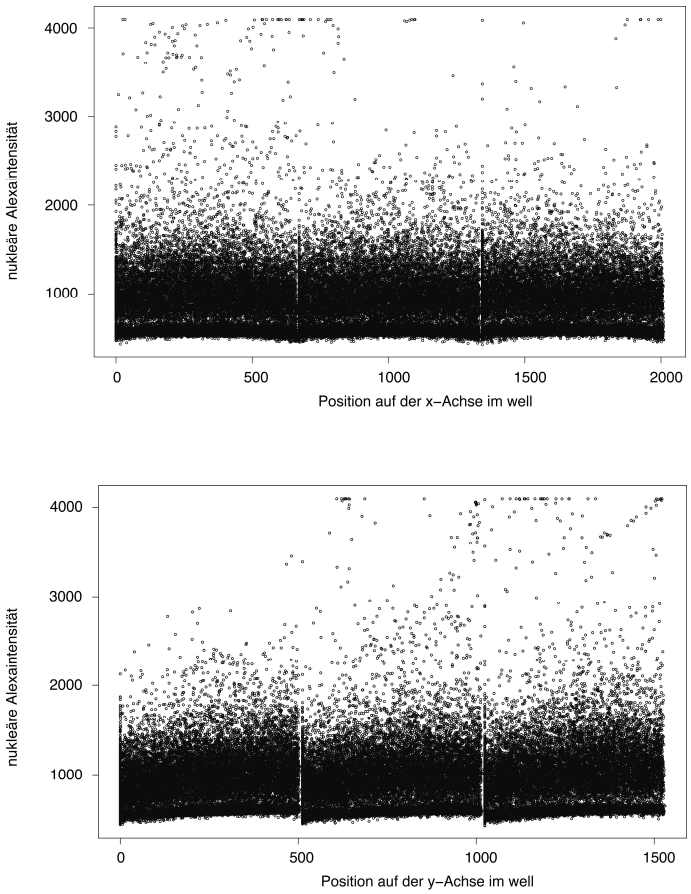


Abb. 63: Scatterplot der Position der Zelle im Bild gegen die Intensität des Alexa-Farbstoffes. Bei dieser Abbildung wird für jede Zelle die Position innerhalb der *wells* – im oberen Bild in der x-Achse, im unteren Bild in der y-Achse - gegen die Intensität aufgetragen. Hiermit soll überprüft werden, ob noch Varianzen auftreten, die durch den Einfallswinkel des Lichtes bedingt sind. Bei dieser Auswertung wurden insgesamt 3x3 Bilder aufgenommen. Ein Bogen stellt die Werte der 3 übereinander liegenden (x-Achse) bzw. der 3 nebeneinander liegenden (y-Achse) Bilder dar. Die Intensitäten zeigen eine weitgehend x- bzw. y-Achsen unabhängige Verteilung der Werte.

Testen der zeitlichen Stabilität des Screens

Die zeitliche Stabilität des Screens wurde getestet, indem die Platten an zwei aufeinanderfolgenden Tagen gemessen wurden. Die Ergebnisse der beiden Messungen wurden im Scatterplot gegeneinander aufgetragen.

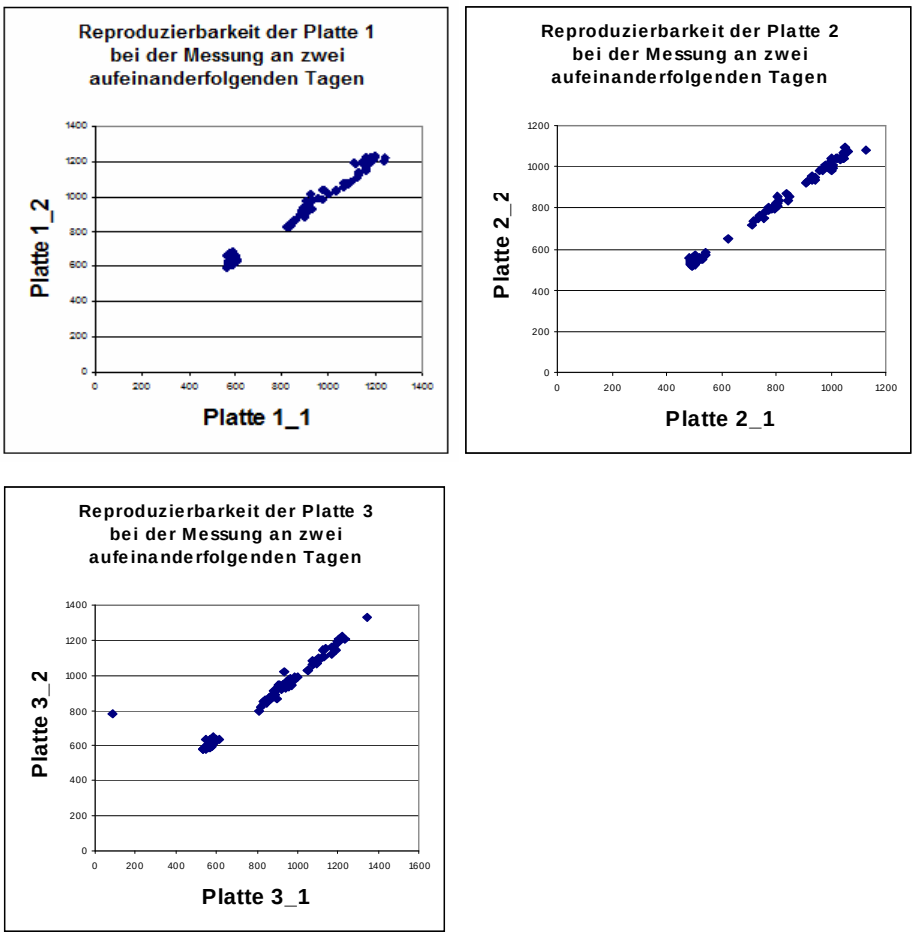


Abb. 64: Scatterplot der an aufeinander folgenden Tagen gemessen Intensitätswerte einer Platte. Die Intensitätswerte, die an aufeinanderfolgenden Tagen für jede Platte gemessen wurden, werden im Scatterplot gegeneinander aufgetragen. Alle drei Platten zeigen auch nach einem Tag eine hohe Reproduzierbarkeit.

Die Scatterplots zeigen für alle drei Platten eine hohe Reproduzierbarkeit. Die Platten sind auch noch nach einem Tag stabil.

Der Vorversuch hat gezeigt, dass die gewählten Versuchsparameter für die Durchführung eines Screens geeignet sind. Es liegt ein Messfenster vor, welches ausreicht, um shRNA Effekte messen zu können. Auch wenn reihenabhängige Effekte auftreten, sind diese gering und beeinflussen die Datenqualität nur wenig. Es treten deutliche Randeffekte auf, die sich vor allem in der Zellzahl bemerkbar machen. Da die Werte innerhalb der *wells* sehr variieren, muss darauf geachtet werden, dass eine ausreichend hohe Zellzahl pro *well* vorliegt, um stabile Messungen zu erhalten.

Es wurden kaum Unterschiede zwischen den Messungen am ersten und am zweiten Tag beobachtet. Der Screen ist somit für ein Zeitfenster von einem Tag, welches für die Messung dieses Screens ausreicht, stabil.

6.3.4 Statistische Auswertung eines shRNA Kinase Screens

Da die Versuchsbedingungen für den Screen geeignet sind, wurde der Screen wie oben beschrieben durchgeführt. Ziel des Screens ist die Untersuchung des Einflusses der verschiedenen Kinasen nach dem Ausschalten mittels einer shRNA auf die Stabilität von c-Myc. Der Screen wurde 3 Mal wiederholt, so liegen Daten dieser 3 Screens mit jeweils 8 x 96-*well*-Platten vor.

Qualitätsparameter und diagnostische Plots

Zunächst wurden die einzelnen Platten bzw. der gesamte Versuch hinsichtlich der Qualität bewertet. Anhand der negativen und positiven Kontrollen wurden die in Tabelle 16 gezeigten Qualitätsparameter berechnet.

Die meisten Z´-Faktoren liegen im Bereich von -0,1 bis 0,7, d.h. von tolerabel bis gut. Vier Platten (Screen 1 – Platte 2, Screen 3 - Platte 2, Screen 3 - Platte 4 und Screen 3 - Platte 7) zeigen deutlich niedrigere Z´-Faktoren bzw. Signalfenster. Bei den drei ersten Platten hängt dies mit einer Erhöhung der Standardabweichung der negativen Kontrollen zusammen und bei der vierten Platte mit der Erhöhung der Standardabweichung der positiven Kontrollen und einer relativ geringen medianen Intensität der positiven Kontrollen. Die Signal-Hintergrund-Verhältnisse sind bei Screen 1 und 2 mit Werten unter 2 recht gering. Die Signal-Rausch-Verhältnisse zeigen dagegen gute Werte.

Tab. 16: Qualitätsparameter des shRNA Kinase Screens (3 Wiederholungsmessungen)

Screen 1

	Platte_1	Platte_2	Platte_3	Platte_4	Platte_5	Platte_6	Platte_7	Platte_8
Median der positiven Kontrollen	888,24	923,39	1044,77	1040,81	894,68	798,41	850,5	841,45
Standardabweichung der pos. Kontrollen	69,61	21,2	65,08	71,78	89,48	23,5	63,27	32,2
Median der negativen Kontrollen	602,13	685,79	693,92	728,01	589,3	582,42	563,03	582,87
Standardabweichung der neg. Kontrollen	26,29	113,37	15,2	13,45	14,63	9,42	8,61	6,09
Z´-Faktor (median)	-0,01	-0,7	0,31	0,18	-0,02	0,54	0,25	0,56
Signalfenster (median)	-0,02	-7,25	1,45	0,78	-0,08	4,47	1,14	6,1
maximale Signal-Rausch-Verhältnis	12,76	43,56	16,05	14,5	10	33,98	13,44	26,13
minimale Signal-Rausch-Verhältnis	22,9	6,05	45,64	54,12	40,29	61,82	65,43	95,64
Signal-Hintergrund-Verhältnis	1,48	1,35	1,51	1,43	1,52	1,37	1,51	1,44

Screen 2

	Platte_1	Platte_2	Platte_3	Platte_4	Platte_5	Platte_6	Platte_7	Platte_8
Median der positiven Kontrollen	802,27	879,26	899,53	963,24	779,1	666,82	881,98	770,65
Standardabweichung der pos. Kontrollen	37,95	126,71	75,63	60,21	32,3	76,62	58,4	68,36
Median der negativen Kontrollen	467,8	447,53	478,69	481,98	416,57	431,39	481,85	436,9
Standardabweichung der neg. Kontrollen	13,74	3,67	65,97	11,02	10,07	6,93	4,38	13,53
Z´-Faktor (median)	0,54	0,09	-0,01	0,56	0,65	-0,06	0,53	0,26
Signalfenster (median)	4,84	0,29	-0,05	4,17	5,45	-0,18	7,93	1,36
maximale Signal-Rausch-Verhältnis	21,14	6,94	11,89	16	24,12	8,7	15,1	11,27
minimale Signal-Rausch-Verhältnis	34,05	122,05	7,26	43,72	41,35	62,26	110,08	32,28
Signal-Hintergrund-Verhältnis	1,71	1,96	1,88	2	1,87	1,55	1,83	1,76

Screen 3

	Platte_1	Platte_2	Platte_3	Platte_4	Platte_5	Platte_6	Platte_7	Platte_8
Median der positiven Kontrollen	1145,13	1055,83	1030,72	740,63	1082,53	1036,21	787,65	1023,98
Standardabweichung der pos. Kontrollen	88,23	98,56	133,8	166,18	89,26	53,24	121,74	108,8
Median der negativen Kontrollen	490,67	505,92	460,75	754,09	510,1	494,82	466,77	511,57
Standardabweichung der neg. Kontrollen	19,6	154,28	18,05	141,25	10,79	15,64	8,98	8,57
Z´-Faktor (median)	0,51	-0,38	0,2	-67,55	0,48	0,62	-0,22	0,31
Signalfenster (median)	3,91	-2,16	0,91	-5,05	3,09	6,34	-0,54	1,44
maximale Signal-Rausch-Verhältnis	12,98	10,71	7,7	4,46	12,13	19,46	6,47	9,41
minimale Signal-Rausch-Verhältnis	25,04	3,28	25,53	5,34	47,26	31,64	51,96	59,69
Signal-Hintergrund-Verhältnis	2,33	2,09	2,24	0,98	2,12	2,09	1,69	2

Um die Verteilung der nukleären Alexaintensität der Kontrollen und der mit shRNA behandelten Zellen zu überprüfen, wurden in Abbildung 65 die Messwerte der negativen und positiven Kontrollen bzw. der shRNAs aller Platten des Screens als Boxplots aufgetragen.

Die Verteilung zeigt, dass die negativen Kontrollen der Platte 2 des Screens 1 und der Platte 2 des Screens 3 durch einen Ausreißerwert eine erhöhte Standardabweichung haben, während die Werte der Platte 4 des Screens 3 insgesamt eine größere Spannbreite haben als die negativen Kontrollen der anderen Platten. Zudem liegen die negativen Kontrollen von Platte 4 des Screens 3 auch im Intensitätsbereich deutlich höher als die negative Kontrollen der Platten des gleichen Screens. Dagegen sind die positiven Kontrollen dieser Platte niedriger als die positiven Kontrollen der anderen Platten. Auch die shRNA-Werte der Platte 4 des Screens 3 zeigen eine leicht höhere Streuung der Daten als bei den benachbarten Screens, aber die mediane Intensitätshöhe entspricht der der benachbarten Platten. Somit werden hier vermutlich nur die Kontrollen stärker beeinflusst und nicht die gesamte Platte. Insgesamt entspricht die Veränderung des Intensitätsbereichs der Kontrollen für die verschiedenen Platten der 3 Screendurchgänge dem Intensitätsbereich der shRNA. Nur bei Screen 3 zeigen die Kontrollen einen veränderten Verlauf. Dies zeigt, dass man nicht ausschließlich von den Kontrollen, und somit von Z´-Faktor, auf die Qualität der Platten schließen kann.

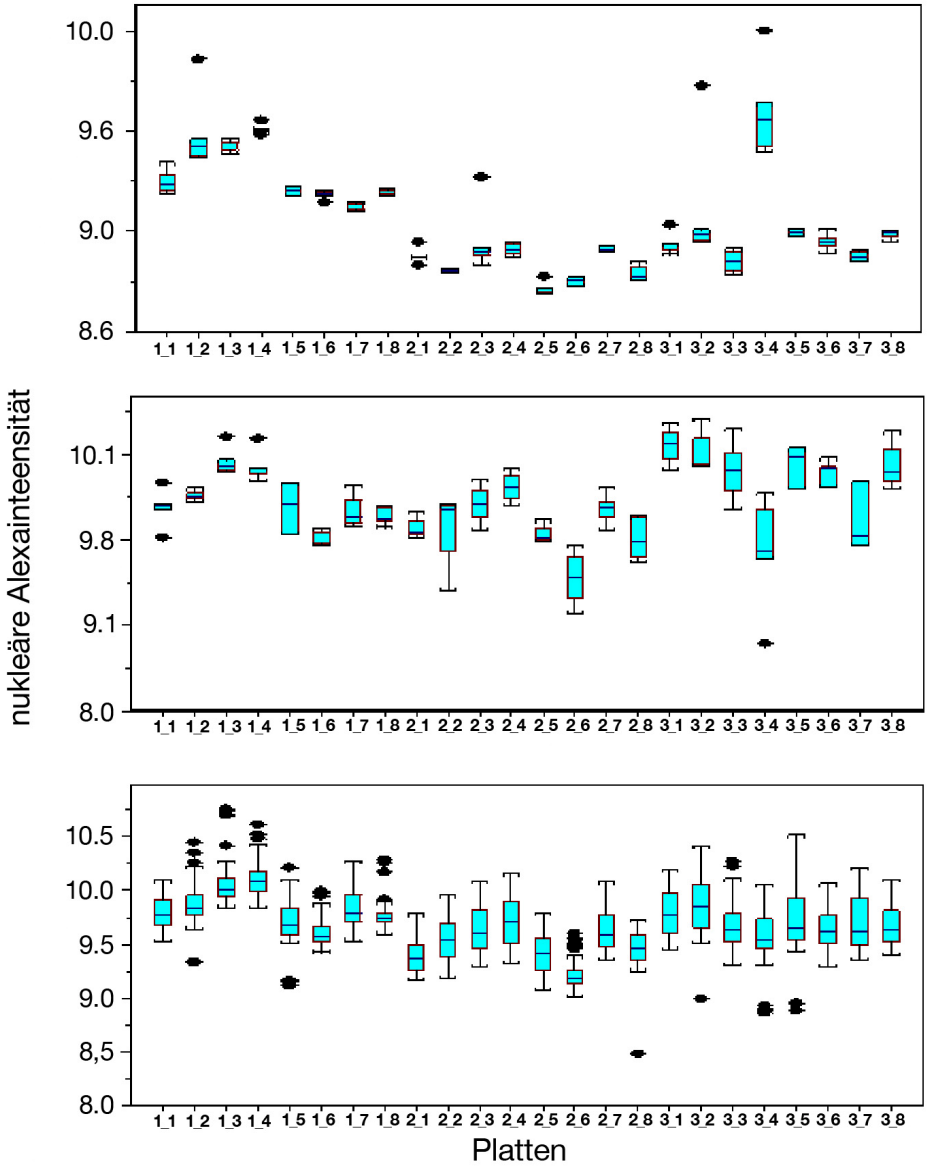


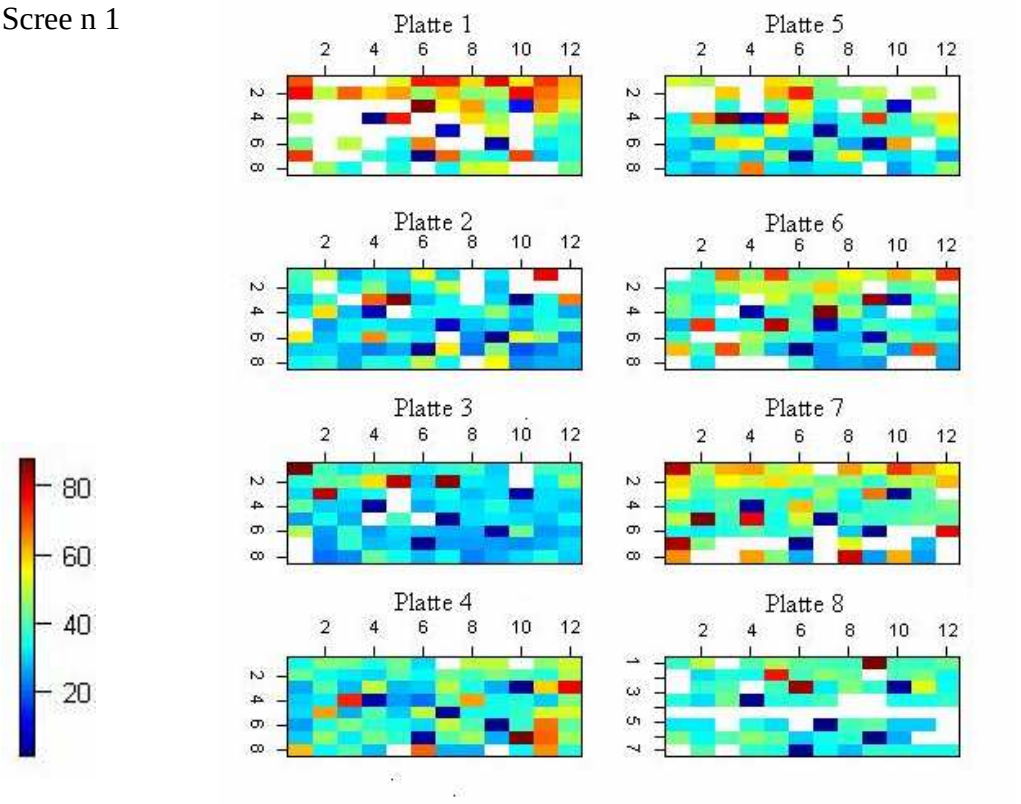
Abb. 65: Boxplots der Intensitäten des nukleären Alexa-Farbstoffes (nach log_2-Transformation), der (a) negative und (b) positiven Kontrollen bzw. (c) der *wells* mit shRNA Behandlung für jede Platte des Screens. Mit Boxplots wird die Häufigkeitsverteilung der Daten dargestellt. Die untere und obere Linie der einzelnen Boxplots stellen das erste und dritte Quartil, die mittlere Linie den Median des jeweiligen Mikroarrays dar. Das Ende der Linien über und unterhalb der Box zeigt das Maximum beziehungsweise das Minimum einer Verteilung an. Werte, die mehr als das 1,5-fache des Interquartil-abstands vom Median abweichen, werden als Ausreißer bewertet und werden als einzelne Datenpunkte dargestellt. Die Boxplots zeigen, dass die Platte 2 des Screen 1 und die Platte 2 des Screen 3 jeweils eine Ausreißerwert innerhalb der negativen Kontrollen haben. Die Platte 4 des Screen 3 hat für die negativen Kontrollen eine größere Spannbreite und liegt bei höheren Intensitäten als die anderen Platten des Versuches. Die positiven Kontrollen dieser Platte sind dagegen niedriger als die positiven Kontrollen der anderen Platten. Auch die shRNA-Werte der Platte 4 des Screen 3 zeigen eine leicht höhere Streuung der Daten als bei den benachbarten Screens, aber die mediane Intensitätshöhe entspricht der der benachbarten Platten. Insgesamt entspricht die Veränderung des Intensitätsbereichs der Kontrollen für die verschiedenen Platten der 3 Screendurchgänge dem Intensitätsbereich der shRNA. Nur bei Screen 3 zeigen die Kontrollen einen veränderten Verlauf.

Bildplots der Platten

In Abbildung 66 sind die nukleären Alexaintensitäten als Bildplots für alle Platten dargestellt. Mit dieser Abbildung sollten die einzelnen Platten auf mögliche Reihen-, Spalten- bzw. Eckeneffekte untersucht werden.

Die Bilder zeigen Eckeneffekte der Platten auf. Auf einigen der Platten nimmt die Intensität der *wells* in der rechten unteren Ecke ab. Aussagen über mögliche systematische Reihen- bzw. Spalteneffekte können anhand der Bildplots nicht gemacht werden.

Scree n 1



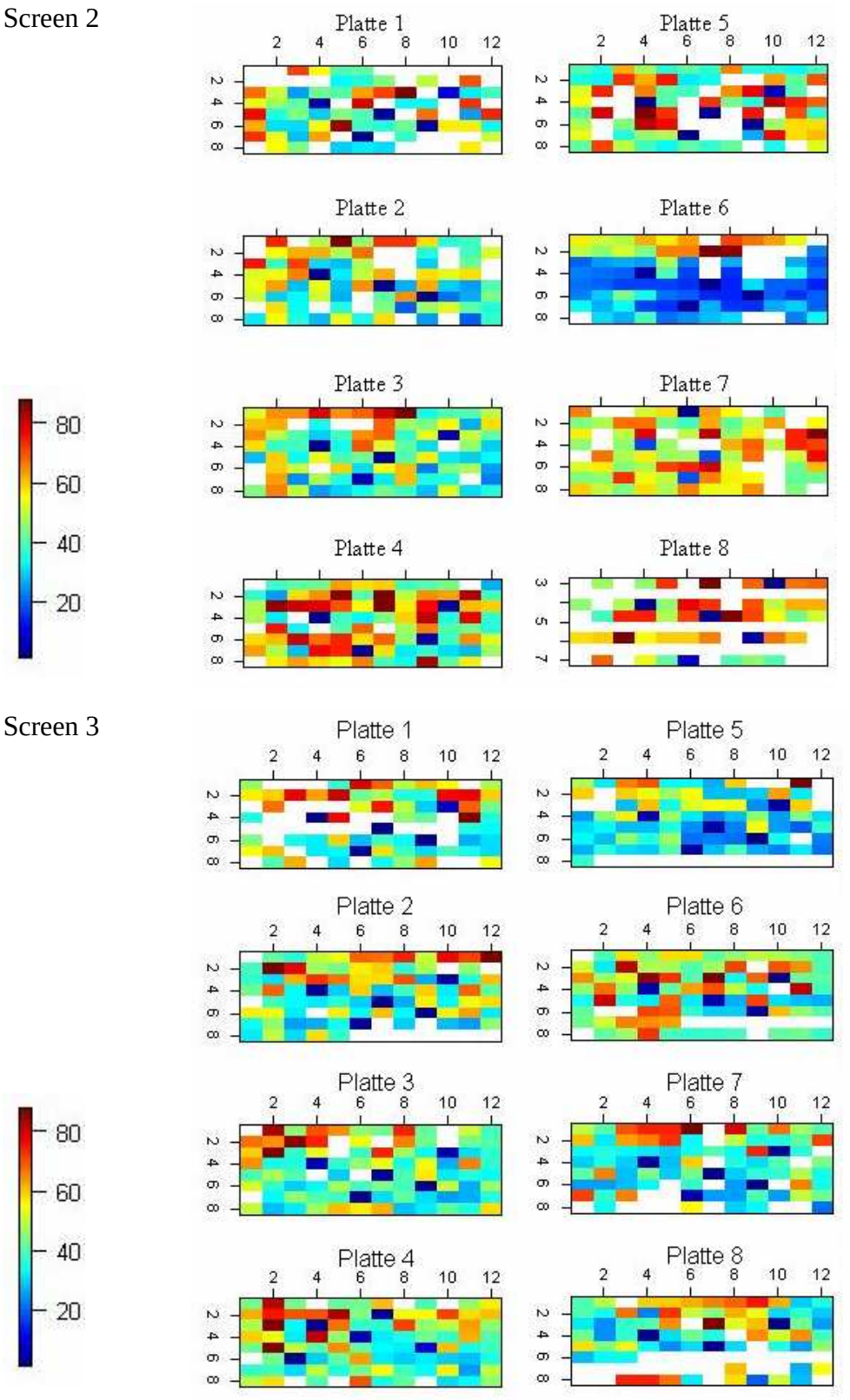


Abb. 66: Bildplots der mittleren nukleären Alexaintensitäten der jeweils 8 Platten der drei Screens Mit den Bildplots soll geprüft werden, ob positionsabhängige Effekte des nukleären Alexa-Farbstoffes auftreten. Hierzu wird die mittlere Intensität des nukleären Alexafarbstoffes für die einzelnen *wells* entsprechend ihrer Position auf der Platte farblich kodiert dargestellt. Weiße Felder sind *wells*, die aufgrund einer niedrigen Zellzahl (< 50 Zellen) von der Auswertung ausgeschlossen wurden. Die Bilder zeigen Eckeneffekte der Platten auf. Auf einigen der Platten nimmt die Intensität der *wells* in der rechten unteren Ecke ab, was anhand der zunehmenden Blaufärbung in den rechten unteren Ecken zu sehen ist.

Um mögliche Reihen- bzw. Spalteneffekte abschätzen zu können, wurden vom gesamten Versuch, d.h. allen drei Screens, die Intensitäten der *wells* in Boxplots für jede Reihe bzw. Spalte dargestellt (s. Abbildung 67).

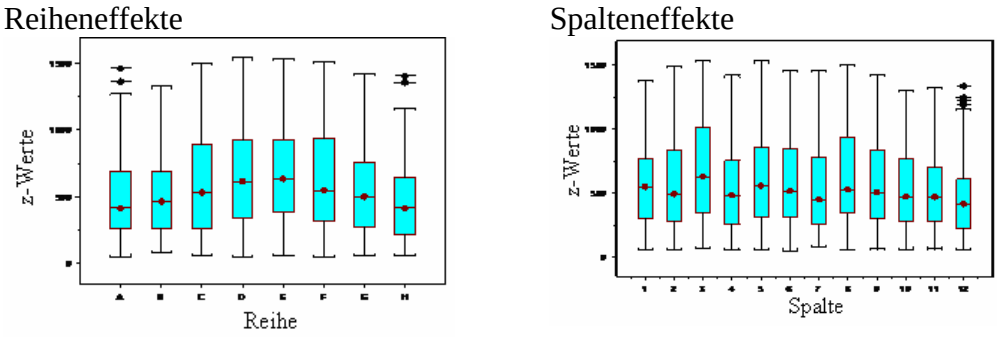


Abb. 67: Boxplots der nukleären Alexaintensitäten der Reihen (links) bzw. Spalten (rechts) aller Platten. Die Boxplots stellen die Verteilung der mittleren nukleären Alexaintensitäten der *wells* in Abhängigkeit von der Reihe bzw. Spalte dar. Die Intensitäten in den äußeren Reihen (A, B, G, H) sind niedriger als die Intensitäten in den mittleren Reihen. Bei den Spalten sieht man eine leichte Abnahme der Intensitäten in den Spalten 10-12.

Die Intensitäten in den äußeren Reihen (A, B, G, H) sind niedriger als die Intensitäten in den mittleren Reihen. Bei den Spalten sieht man eine leichte Abnahme der Intensitäten in den Spalten 10-12.

Logarithmische Transformation der Daten

Im nächsten Schritt wurden die Daten auf eine $\log_2$-Skala transformiert, um so eine Normalverteilung der Daten zu generieren, die Voraussetzung für verschiedene statistische Tests.

Normalisierung der Daten

Falls systematische Unterschiede zwischen den Platten der Screens vorliegen, z.B. infolge von Antikörpern verschiedener Chargen, unterschiedliche Belichtungszeiten usw., müssen die Daten normalisiert werden, um die Ergebnisse miteinander vergleichen zu können. Mit den folgenden diagnostischen Plots werden die Daten auf die Notwendigkeit der Normalisierung überprüft.

Abbildung 68 zeigt einen Scatterplot, der auf der x-Achse die laufenden Nummern, d.h. die *wells* aller Platten in der Reihenfolge in der sie gemessen wurden, gegen die Rohdaten der nukleären Alexaintensität darstellt.

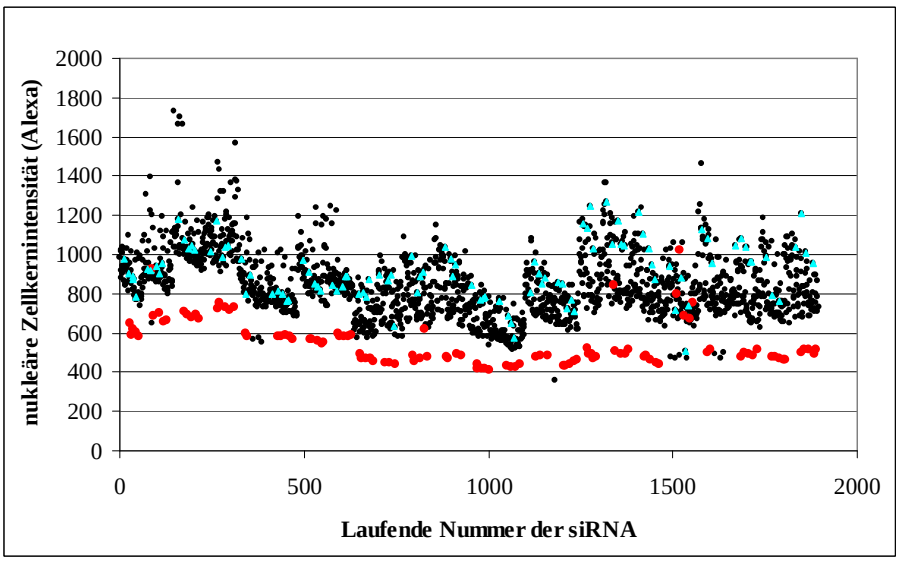


Abb. 68: Scatterplot der *wells* mit den verschiedenen shRNAs gegen die nukleäre Alexaintensität. Die *wells* sind in der Reihenfolge aufgetragen in der sie ausgewertet wurden, d.h. die Auftragung erfolgt von Screen 1 Platte 1 bis zu Screen 3 Platte 8. Innerhalb einer Platte sind die *wells* reihenweise aufgetragen. Die positiven Kontrollen sind als gelbe Dreiecke dargestellt, die negativen Kontrollen als rote Kreise. Die Intensitätsbereiche weisen für die unterschiedlichen Platten deutliche „Sprünge“ auf, so dass eine Normalisierung notwendig ist.

Im Scatterplot sieht man deutlich „Sprünge“ der Intensitätsbereiche, sowohl bei den Screens untereinander als auch bei den einzelnen Platten innerhalb eines Screens. Dies macht die Notwendigkeit eines Normalisierungsschrittes deutlich.
Weiterhin zeigt der Plot, dass die positiven Kontrollen in Screen 1 innerhalb der Daten-„Wolken“ liegen, während sie in Screen 2 und 3 eher im oberen Bereich der Daten-„Wolken“ liegen. In Screen 2 und 3 wurde eine andere Produktionscharge des *shScramble*-Vektors verwendet.

Ein weiterer systematischer Fehler, der mittels eines Normalisierungsschrittes korrigiert werden muss, ist eine zellzahlabhängige Varianz der Daten. Bei der Betrachtung der Mikroskopbilder erschienen die Intensitäten des Alexa-Farbstoffes in Bildern mit wenigen Zellen höher als in Bildern mit vielen Zellen. Um zu überprüfen, ob es sich hierbei um einen systematischen Effekt handelt, wurde die Anzahl der gemessenen Zellen innerhalb der einzelnen *wells* gegen die mittleren nukleären Alexaintensitäten des *wells* im Scatterplot dargestellt (s. Abbildung 69).

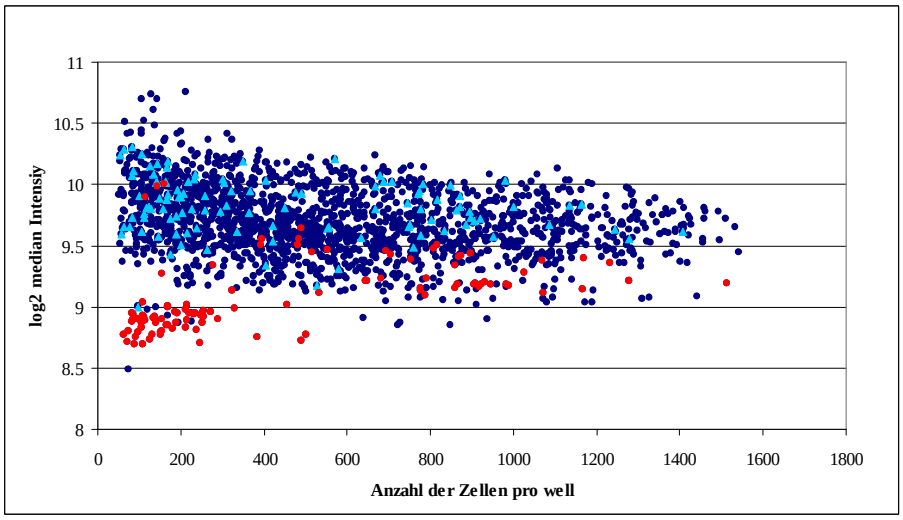
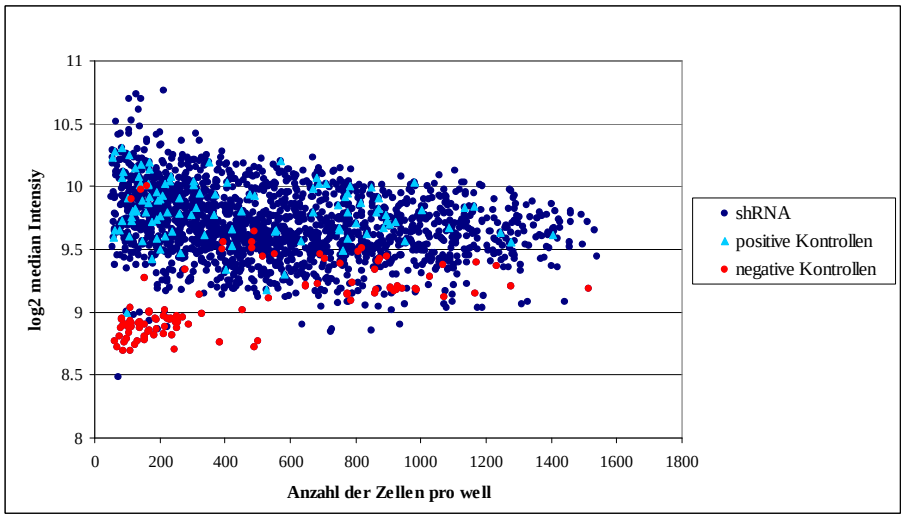


Abb. 69: Scatterplot der Anzahl der Zellen eines *wells* gegen die nukleären Alexaintensitäten In diesem Plot soll eine systematische Abhängigkeit der nukleären Alexaintensität von der Zellzahl innerhalb der *wells* überprüft werden. Hierzu wird in einem Scatterplot die Anzahl der Zellen in einem *well* (x-Achse) gegen die mittlere nukleäre Intensität in dem *well* aufgetragen. Die Abbildung zeigt, dass mit zunehmender Zellzahl die mittlere Intensität abnimmt. Dies gilt nur für Zellen, die mit dem ersten Antikörper behandelt wurden, da bei den negativen Kontrollen die Intensität mit zunehmender Zellzahl zunimmt.

Hier ist eine deutliche Abhängigkeit der Intensität von der Zellzahl zu sehen. Je niedriger die Zellzahl, desto höher ist die Intensität. Dies gilt jedoch nur für die Proben, die mit dem ersten Antikörper behandelt wurden. Die negativen Kontrollen, bei denen der erste Antikörper weggelassen wurde, zeigen eher einen umgekehrten Effekt, ihre Intensität steigt mit zunehmender Zellzahl. Somit scheint es sich hierbei um einen technischen Effekt zu handeln, der nur im Zusammenhang der Verwendung des 1. Antikörpers auftritt. Die Korrektur dieses Effektes erfolgte mit einer Lowess-Normalisierung der einzelnen Platten, wie sie im Rahmen der Datenauswertung der Mikroarrayversuche beschrieben wurde (s. Kapitel 2.2.6). In Abbildung 70 wird die Zellzahl-abhängige Verteilung der Intensitäten vor (a) und nach (b) der Lowess-Normalisierung dargestellt. Die Zellzahl-abhängige Effekte wurden durch die Lowess-Normalisierung aufgehoben. Die Variabilität der negativen Kontrollen wurde durch den Normalisierungsschritt erhöht, was zu erwarten war, da bei ihnen der Zellzahl-abhängigen Effekt nicht zu beobachten war. Da der Normalisierungsschritt zu einer Veränderung des Skalenniveaus führt, wurden die negativen Kontrollen dennoch mit in den Normalisierungsschritt einbezogen.

a)



b)

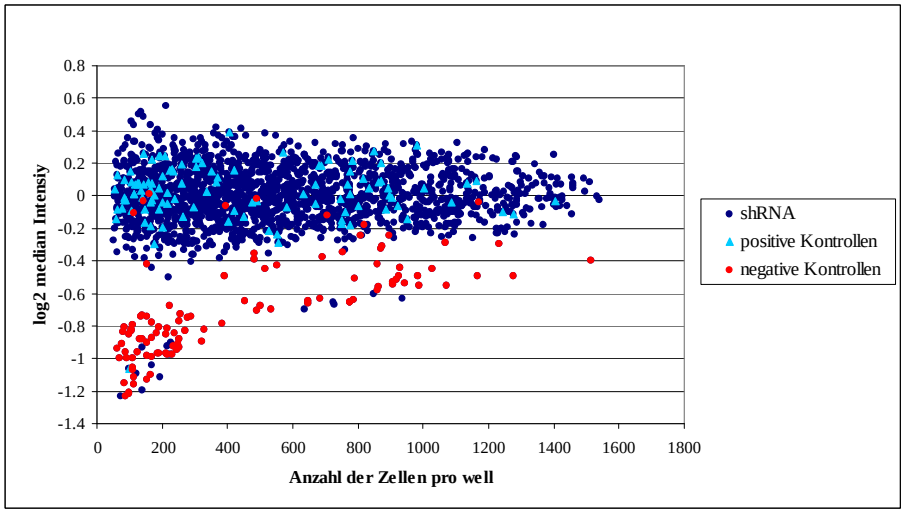
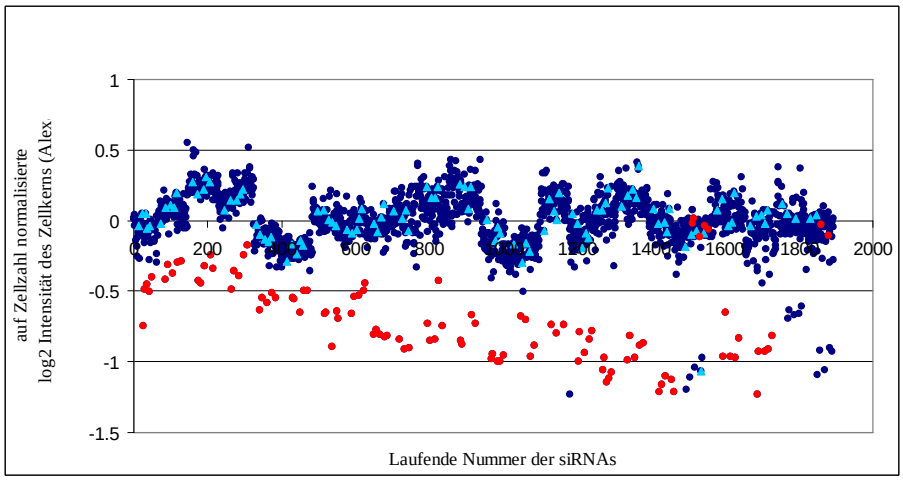


Abb. 70: Scatterplot der Anzahl der Zellen eines *wells* gegen die nukleären Alexaintensitäten des *wells* a) vor der Lowess Normalisierung b) nach der Lowess Normalisierung.

Um die Plattenunterschiede auszugleichen, wurde eine z-Wert Transformation durchgeführt (s. Abbildung 71). Es treten zwar Reihenunterschiede auf, die durch eine B-Wert Transformation korrigiert werden könnten, diese sind jedoch sehr gering, so dass eine z-Wert Transformation aufgrund der höheren statistischen Aussagekraft einer B-Wert Transformation vorzuziehen ist.

a)



b)

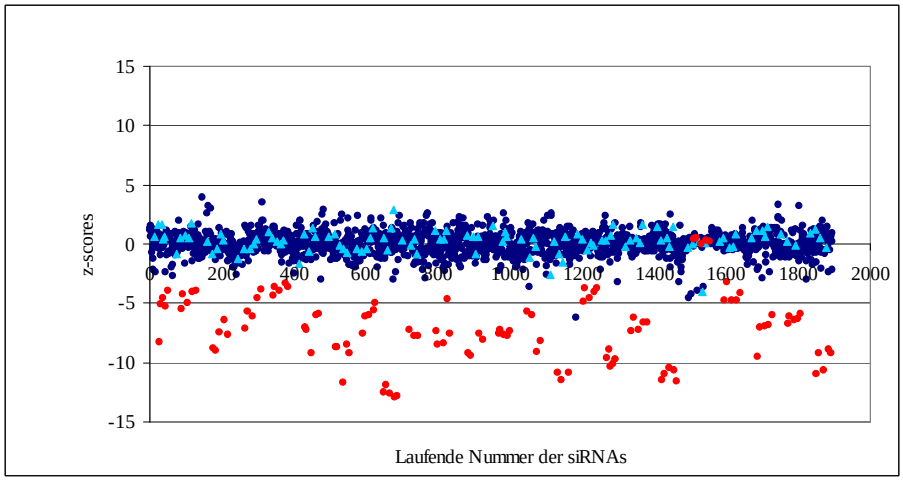


Abb. 71: Scatterplot der shRNA gegen (Lowess-) normalisierte nukleären Alexaintensitäten (a) und gegen (Lowess-) normalisierte und z-Wert transformierte nukleären Alexaintensitäten (b) Die shRNAs sind in der Reihenfolge aufgetragen in der sie ausgewertet wurden, d.h. die Auftragung erfolgt von Screen 1 Platte 1 bis zu Screen 3 Platte 8. Innerhalb einer Platte sind die shRNAs reihenweise aufgetragen.

Reproduzierbarkeit der drei Screens

Nach der Normalisierung der Platten wurde die Reproduzierbarkeit der Ergebnisse der drei Screens betrachtet. Hierzu wurde eine Scatterplot-Matrix der 3 Screens erstellt bzw. der Korrelationskoeffizient berechnet.

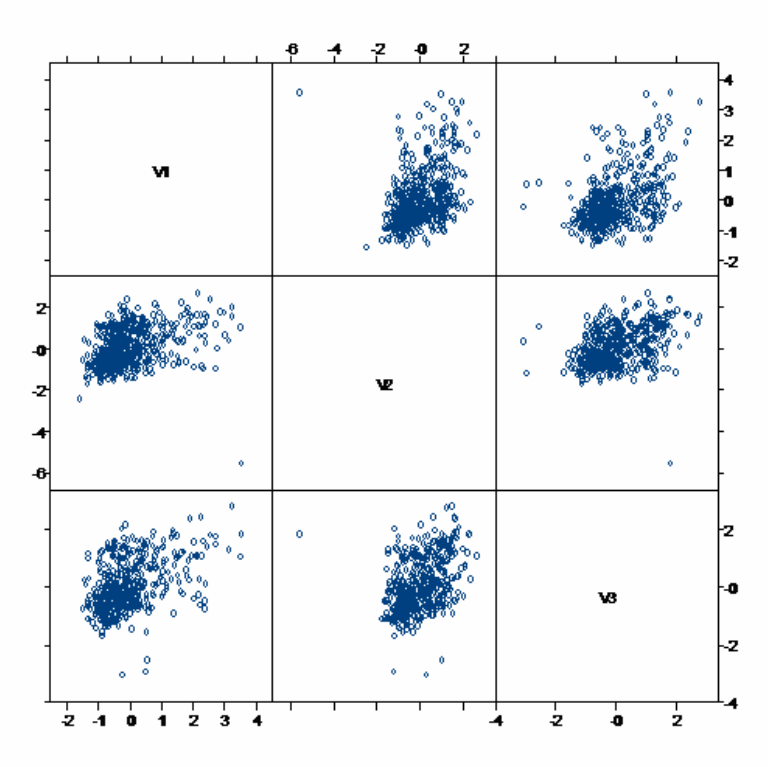


Abb. 72: Scatterplot-Matrix der z-Wertes der drei Wiederholungsscreens

Tab. 17: Korrelationskoeffizienten der drei Wiederholungsscreens

	Screen 1	Screen 2	Screen 3
Screen 1	1	0.31	0.45
Screen 2	0.31	1	0.37
Screen 3	0.45	0.37	1

Die Reproduzierbarkeit von Screen 1 und Screen 3 ist höher im Vergleich zu Screen 2. Da Screen 1 und 3 die gleiche Plattenbelegung haben, ist es möglich, dass die geringere Korrelation mit Screen 2 auf räumliche Effekte auf den Platten zurückzuführen ist. Das System erweist sich dennoch als so stabil, dass die Ergebnisse reproduzierbar sind.

Selektion von *hits*

Zur Selektion von *hits* wurde die *Volcanoplot*-Methode gewählt. Hierzu wurden zunächst die Mittelwerte und die Standardabweichungen der z-Werte der drei Screens berechnet. Die Mittelwerte der z-Werte wurden dann im Scatterplot auf der y-Achse gegen die laufende Nummer der shRNAs auf der x-Achse aufgetragen (s. Abbildung 73).

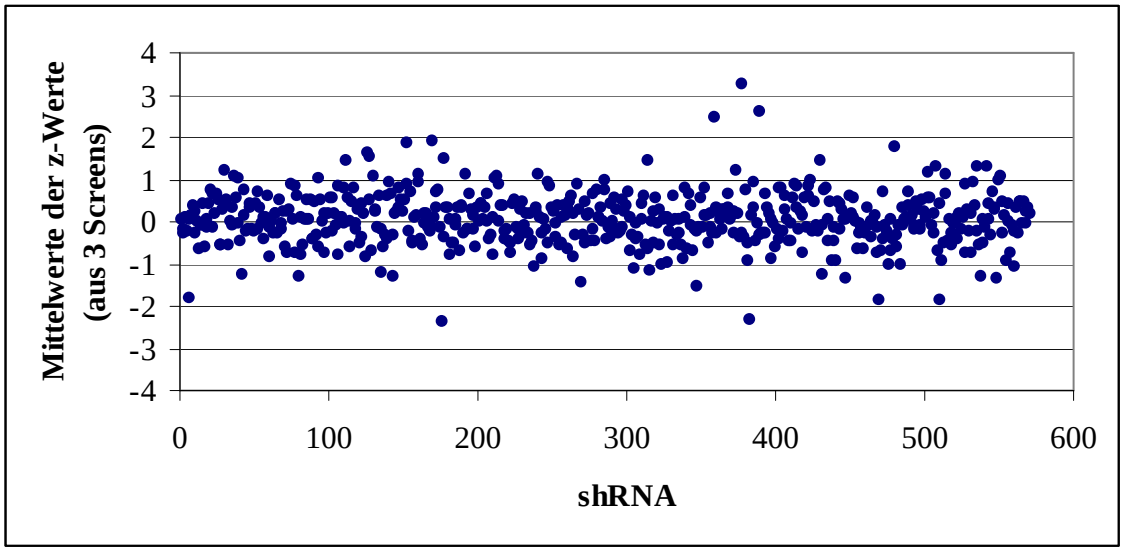


Abb. 73: Scatterplot der shRNAs gegen normalisierte und z-Wert transformierte nukleäre Alexaintensitäten Hier werden die shRNAs gegen die als z-Wert transformierte nukleäre Alexaintensität aufgetragen**.**

Anhand dieser Abbildung wurde für die Auswahl von *hits* ein Schwellenwert für den z-Wert von 1 gewählt. Somit werden alle shRNA-behandelte *wells*, deren Wert sich mehr als eine Standardabweichung über oder unter dem Median, der über alle *wells* der Platte (ohne Kontrolle) berechnet wurde, liegen, für die Selektion von *hits* in Betracht gezogen. Um auch die Streuung der Daten mit in die Selektion einzubeziehen, wurde anschließend die t-Statistik für jede shRNA berechnet und ein *Volcanoplot* erstellt (s. Abbildung 74). Die Auswahl von *hits* erfolgte anhand eines mittleren absoluten z-Wertes von 1 und einer absoluten t-Statistik von 1.96. Anhand dieser Auswahlkriterien wurden 11 shRNAs als *hits* ausgewählt s. Tabelle 18, von denen 8 shRNA zu einer Erhöhung und 3 shRNAs zu einer Reduzierung des c-Myc-Levels in der Zelle geführt haben.

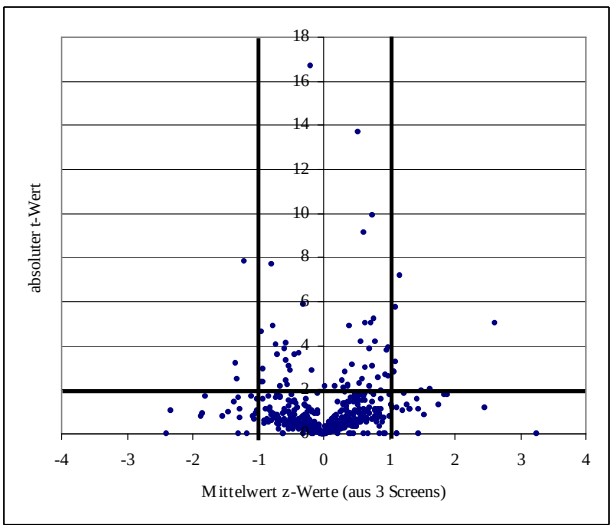


Abb. 74: *Volcanoplot* zur Selektion von shRNA, die die Stabilität von c-myc beeinflussen. Im *Volcanoplot* wird der Mittelwertes der z-Werte der 3 Screens gegen die entsprechende (absolute) t-Statistik der 3 Screens geplottet. Die Linien stellen die Grenzen zur Selektion dar. shRNAs, die außerhalb der Linien liegen, werden als *hits* selektiert

Tab. 18: Liste von shRNA welches aufgrund der Screening-Ergebnisse als *hits* selektiert wurden

RefSeq ID	Unigene Cluster ID	Name	Symbol	z-score Screen1	z-score Screen2	z-score Screen3	MW z-score	t-test
NM_058241	Hs.591241	Cyclin T2	CCNT2	1.03	0.94	1.3	1.09	5.76
NM_015716	Hs.443417	Misshapen-like kinase 1 (zebrafish)	MINK1	0.82	1.61	2.44	1.63	2.01
NM_015092	Hs.460179	PI-3-kinase-related kinase SMG-1	SMG1	0.8	1.34		1.07	2.78
NM_014791	Hs.184339	Maternal embryonic leucine zipper kinase	MELK	-1.06	-1.36	-1.17	-1.2	-7.85
NM_007118	Hs.130031	Triple functional domain (PTPRF interacting)	TRIO	1.93	0.62	1.96	1.51	1.96
NM_005465	Hs.498292	V-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)	AKT3	0.86		1.33	1.09	3.27
NM_002754	Hs.178695	Mitogen-activated protein kinase 13	MAPK13	2.22	2.4	3.2	2.61	5
NM_000297	Hs.181272	Polycystic kidney disease 2 (autosomal dominant)	PKD2	1.21	0.99	1.31	1.17	7.19
AL122055	Hs.584867	Cell division cycle 2-like 6 (CDK8-like)	CDC2L6	-1.93	-0.99	-1.02	-1.31	-2.46
AF119042	Hs.490287	Tripartite motif-containing 24	TRIM24	-1.66	-1.52	-0.87	-1.35	-3.18
AF076974	Hs.203952	Transformation/transcription domain-associated protein	TRRAP		1.35	0.8	1.08	2.8

6.4 Diskussion

Die Verwendung von RNA-Interferenz als Werkzeug zur funktionellen Analyse von Genen gewinnt in der molekularbiologischen und medizinischen Forschung immer mehr an Bedeutung. In den letzten Jahren wurden Systeme entwickelt, die eine Umsetzung der RNAi-Technologie als Hochdurchsatz-Methode ermöglichen. Hochdurchsatz-Screens generieren sehr große Datenmengen, die zudem häufig eine sehr hohe Variabilität aufweisen. Eine möglichst effiziente Datenanalyse ist somit von großer Bedeutung für den Erfolg eines solchen Assays.

Die Voraussetzungen für das optimale experimentelle Design eines RNAi *Imaging* Screens wurde in Abschnitt 6.2.1. dieses Kapitels beschrieben. In dem hier dargestellten Datensatz wurde die Verteilung der shRNAs in einem der drei Screens im Vergleich zu den anderen beiden Screens verändert. Auch wurden die Kontrollen randomisiert über die Platten verteilt. Die Reproduzierbarkeit der einzelnen Screens untereinander zeigt die Bedeutung der Veränderung der Positionen der shRNAs (s. Abbildung 72) auf. So haben die beiden Screens mit der gleichen Plattenbelegung eine höhere Reproduzierbarkeit im Vergleich zu dem Screen mit der unterschiedlichen Plattenbelegung. Die höhere Reproduzierbarkeit ist vermutlich auf räumliche Effekte auf den Platten zurückzuführen. Um stärker auf die biologischen Effekte zu fokussieren, ist eine unterschiedliche Verteilung der shRNAs in den verschiedenen Screens zu bevorzugen.

Der Datensatz zeigt, wie erwartet, eine hohe Variabilität (s. Abbildung 71a). Die experimentelle Durchführung eines Screens basiert auf einer Vielzahl von aufeinander folgenden Schritten. Auch wenn bei der Durchführung eines Screens auf die Einhaltung von standardisierten Versuchsbedingungen genau geachtet wird, ist es dennoch sehr schwierig konstante Bedingungen aufrecht zu erhalten und systematische Fehler vollständig zu vermeiden. Mittels der durchgeführten Qualitätskontrollen können systematische Fehler innerhalb des Datensatzes detektiert werden. So zeigte der Scatterplot in Abbildung 71a deutliche „Intensitätssprünge“ zwischen den verschiedenen Screens aber auch zwischen den verschiedenen Platten eines Screens. Ein weiterer systematische Fehler sind die beobachteten Reiheneffekte.

Die Relation der Intensitäten der positiven Kontrollen in den 3 Screens im Vergleich zu den Intensitäten der mit shRNAs behandelten *wells* demonstriert, welche Auswirkung der Wechsel der Reagenzien-Charge auf das Ergebnis haben kann.

Diese systematischen Fehler können zwar weitgehend mittels statistischer Methoden korrigiert werden, aber mit jeder – wenn auch notwendigen – Korrektur wird in die Datenverteilung „eingegriffen“ und die Aussagekraft der Daten auch verändert.

Die Zellen innerhalb eines *wells* zeigen ebenfalls eine hohe Streuung der Werte. Hier zeigt sich ein generelles Problem bei der Durchführung von Experimenten mit Zellpopulationen. Auch wenn die externen Parameter genau kontrolliert werden und die Zellpopulation gleich behandelt wird, hat dennoch jede Zelle immer noch ihren „eigenen“ Zustand, der von einer Vielzahl von Parametern abhängt. Es wird versucht, die „mittlere Zelle“ zu berechnen, wozu jedoch eine ausreichende Zahl an Zellen notwendig ist. Im Vorversuch wird gezeigt, dass die *wells* mit der gleichen Behandlung – auch wenn innerhalb der *wells* eine hohe Streuung vorliegt – gut reproduzierbare Intensitäten zeigen. Dies deutet darauf hin, dass diese „mittlere“ Zelle bei den Versuchen gut getroffen wird, aber auch die Notwendigkeit mit einer ausreichend großen Zellpopulation zu arbeiten.

Es zeigt sich, dass die Verwendung von Kontroll-*wells* und die Durchführung von Qualitätskontrollen für die Bewertung eines Versuches eine wichtige Rolle spielen. Nur so kann die Funktionalität eines Screens bewertet und systematische Fehler detektiert werden. Anhand des unterschiedlichen Verlaufes der positiven Kontrollen in den drei Screens werden jedoch auch die Grenzen der Kontrollen aufgezeigt. Würden diese Werte zur Normalisierung der Daten verwendet, hätte dies, anstelle einer Verbesserung der Vergleichbarkeit der Platten, die Erhöhung der Unterschiede zur Folge. Bei der Selektion von shRNAs, die die c-myc Stabilität erhöhen, wären die shRNAs des Screens 1 überrepräsentiert im Vergleich zu Screen 2 und 3, während bei den shRNAs, die die c-myc Stabilität verringern, die shRNAs des Screen 1 unterrepräsentiert wären. Die Kontrollen sollten generell, aufgrund ihrer hohen Fehleranfälligkeit, nur zur Normalisierung eingesetzt werden, wenn andere Normalisierungsmethoden, auf Grund fehlender Vorraussetzungen nicht angewendet werden können. Falls sie dennoch zur

Normalisierung verwendet werden, sollten sie im Vorfeld auf ihre Verteilung und mögliche Ausreißer untersucht werden.

Ein Vorversuch unter Verwendung von geeigneten Kontrollen sollte vor Beginn jedes Screens durchgeführt werden, um so die Eignung der gewählten Parameter zu testen. In dem hier ausgewerteten Datensatz wurde eine Probe, bei der der 1. Antikörper weg gelassen wurde, als negative Kontrolle gewählt. Durch das Weglassen des 1. Antikörpers gehen jedoch Informationen hinsichtlich der Varianz, die durch die Verwendung dieses Antikörpers bedingt sind, verloren. Diese Kontrolle ist somit zur Bewertung des Messfensters nicht optimal.

Es wurde eine deutliche Abhängigkeit der Intensität von der Zellzahl gezeigt. Je geringer die Zellzahl, desto höher die Intensität des Alexafarbstoffes im Zellkern. Dies gilt jedoch nicht für die Zellpopulationen, die als negative Kontrolle dienten und nicht mit dem ersten Antikörper behandelt wurden. Hier ist eher ein umgekehrter Effekt zu sehen, je höher die Zellzahl, desto höher die Intensität im Zellkern. Hieraus ergibt sich die Frage, ob es sich hierbei um einen biologischen Effekt handelt z.B. dass c-Myc in Abhängigkeit von der Anzahl der Zellen in der Umgebung unterschiedlich reguliert wird oder um einen technischen Effekt, der z.B. in der Natur des ersten Antikörpers liegt. Da dieser Effekt die Streuung der Daten deutlich erhöht, sollte mit weiteren Kontrollversuchen geklärt werden, worauf dieser Effekt beruht. Hierzu könnten z.B. Testscreens mit unterschiedlichen Zellzahlen bzw. steigender Antikörperkonzentration durchgeführt werden.

Mit der automatisierten Bildanalyse ist es möglich, eine Vielzahl von Parametern eines Bildes parallel zu erfassen. So können z.B. für alle mit Fluoreszenzfarbstoffen markierte Zellbestandteile neben der Intensität der Fluoreszenzfarbstoffe die Textur der Oberfläche, die Form des Zellkerns oder die Anzahl der gemessenen Pixel erfasst werden.
Die Berechnung der c-Myc-Level und die Selektion von *hits* erfolgt anhand der nukleären Alexaintensitäten. Die Granularität als weiterer Selektionsparameter kann hier nicht verwendet werden, da in höheren Intensitätsbereichen eine Sättigung der Werte eingetreten ist. Da Intensitätsmessungen, bzw. allgemein formuliert, eindimensional bestimmte Parameter häufig durch unspezifische Faktoren beeinflusst

werden, könnte eine mehrdimensionale Datenanalyse möglicherweise zur effizienteren Selektion von *hits* führen.
Eine besondere Problematik von si-/shRNA Screens ist die unterschiedliche Effizienz der si-/shRNAs. Um diese auszugleichen, werden z.B. bei der hier verwendeten NKI-Bibliothek, 3 shRNA-Vektoren pro Pool eingesetzt. Im Screen wird in allen *wells* die gleiche shRNA-Konzentration eingesetzt. Während diese Konzentration für die Abschaltung einzelner Kinasen nicht ausreicht und ein falsch negatives Ergebnis erzielt wird, werden andere Kinasen aufgrund eines *off-target* Effektes gehemmt und so ein falsch positives Ergebnis erzielt. Ein Phänomen, welches derzeit nur schwer zu kontrollieren ist, aber bei der Interpretation der Daten berücksichtigt werden muss.
Eine weitere Problematik ist die Eigenschaft einiger Kinasen, ihre vollständige Funktionalität zu erzielen, auch wenn sie nur in einer geringen Konzentration innerhalb der Zelle vorliegen.

Im Rahmen dieses Kapitels wurde die Etablierung eines statistischen Datenanalyse-verfahrens zur Auswertung von RNAi *Imaging* Screens beschrieben.

Dieses beinhaltet folgende Schritte:

Experimentelles Design
Qualitätskontrollen
Normalisierung
Selektion von *hits*

Zunächst wurde ein Vorversuch mit einer positiven und einer negativen Kontrolle, bzw. einer Kontrolle, von der ein Intensitätswert zwischen den beiden anderen Kontrollen erwartet wird, durchgeführt. Anhand dieses Versuches konnten sowohl das Messfenster als auch räumliche und zeitliche Effekte gut abgeschätzt werden und daraus geschlossen werden, dass sich die gewählten Bedingungen zur Durchführung eines Screens eignen.
Es wurden drei Wiederholungsscreens durchgeführt. Zwei der Screens haben die gleiche Plattenbelegung, bei dem dritten Screen wurde die Plattenbelegung verändert. Die Kontrollen sind auf allen Platten randomisiert. Die drei Screens weisen eine gute Reproduzierbarkeit auf, die bei den beiden Screens, die die gleiche Plattenbelegung haben jedoch z.T. räumlich bedingt sind.

Anhand der verwendeten Qualitätsparameter und graphischen Darstellungen der Daten konnten systematische Fehler gut dargestellt werden und eine qualitative Bewertung der einzelnen Platten bzw. des Gesamtscreens vorgenommen werden. Die systematischen Fehler konnten durch die Verwendung geeigneter Normalisierungsmethoden d.h. die Lowess-Normalisierung zur Korrektur der Zellzahl-abhängigen Effekte und die z-Transformation zur Angleichung der verschiedenen Platten weitgehend aufgehoben werden. Somit konnten die Daten aller Platten gemeinsam ausgewertet werden und aus der Verteilung des Gesamtdatensatz zusätzliche Informationen gewonnen werden, die z.B. für die Bewertung der Qualität aber auch die Selektion von *hits* von Bedeutung sind. Mittels des *Volcanoplots*, welcher in der Datenanalyse von Mikroarray-Experimenten Verwendung findet, konnte bei der Selektion von *hits*, sowohl die mittlere Veränderung der c-myc-Stabilität in Form des z-Wertes als auch über die t-Statistik die Streuung der Daten einbezogen werden.

7 Zusammenfassung

Der Einsatz von Hochdurchsatz-Methoden ist in der molekularbiologischen Forschung zu einem elementaren Werkzeug zur Aufklärung der komplexen zellulären Vorgänge geworden. Ein wichtiger Aspekt bei der Durchführung von Hochdurchsatz-Methoden ist die Etablierung von geeigneten Algorithmen zur Auswertung der Daten. In der Vergangenheit hat sich gezeigt, dass das Fehlen von standardisierten Qualitätskriterien und Datenanalysemethoden eines der Probleme für die effiziente Implementierung von Hochdurchsatz-Methoden (Kaul 2005) und somit von großer Bedeutung für den Erfolg dieser Versuche ist.

Für die DNA-Chiptechnologie, die bereits seit Anfang der 90er Jahre Anwendung in der Untersuchung der Genexpression findet, sind bereits standardisierte Methoden zur Auswertung der Daten etabliert worden. Ein Überblick über diese Methoden wird in Kapitel 2 gegeben. Es hat sich jedoch gezeigt, dass die Interpretation der Daten im Kontext der biologischen Fragestellung sich trotz effizienter Datenanalyse als schwierig gestaltet. Deshalb wurde stärker darauf fokussiert, bioinformatische Methoden zur funktionellen Interpretation der Daten zu etablieren. In Kapitel 3 werden einige dieser Methoden dargestellt und anschließend auf einen Mikroarray-Datensatz angewendet, mit dem die c-Myc abhängige Genexpression in T-Lymphozyten von transgenen Mäusen untersucht wird. Es konnte gezeigt werden, dass die hier angewendeten Methoden für die funktionelle Interpretation von Genlisten geeignet sind. Da die Methoden zu unterschiedlichen Ergebnissen führen, bzw. auf unterschiedliche Schwerpunkte fokussieren, ist es sinnvoll, mehrere Methoden parallel anzuwenden, um eine möglichst effiziente Interpretation zu erzielen.

Ein Aspekt der funktionellen Analyse ist die Untersuchung von ko-regulierten Genen eines Datensatzes auf eine mögliche Regulation durch einen gemeinsamen Transkriptionsfaktor. Eine Methode zur Detektion signifikant überrepräsentierter cis-regulativer Motive in den Promotersequenzen eines Sets von ko-regulierten Genen wurde im Rahmen der Arbeit etabliert. Diese Methode beinhaltet folgende Schritte:

- Erstellung einer Datenbank mit orthologen Promotorsequenzen (Maus/Mensch)
- Maskierung von repetitiven Sequenzelementen

- Alignment der orthologen Sequenzen
- Untersuchung von konservierten Sequenzbereichen auf Transkriptionsfaktor-Bindungsstellen (TFBS)
- Korrektur der Anzahl der Bindungsstellen auf die Länge der konservierten Promotorsequenzen
- Untersuchung einer Gruppe von ko-regulierten Genen auf die Anreicherung von TFBS im Vergleich zu einem Hintergrundset

Die Methode wurde auf den in Kapitel 3 verwendeten Mikroarray-Datensatzes angewendet. *Position weight matrices* (PWM) mit E-Box-Motiven, die als Bindungssequenz für c-Myc bereits beschrieben wurden, konnten als signifikant überrepräsentiert gefunden werden. Zudem findet sich eine Anreicherung der PWM für den Transkriptionsfaktor YY1, dessen konstitutive Repression durch die Überexpression von c-Myc reduziert wird (Austen, Cerni et al. 1998).

Neben der Regulation der Genexpression z.B. durch Trankriptionsfaktoren gibt es eine Reihe von weiteren Faktoren, die eine wichtige Rolle bei der Regulation der zellulären Prozesse spielen. Hierzu gehören u.a. die miRNAs, die als zelluläre Regulatoren in den letzten Jahren vermehrt in den Fokus der molekularbiologischen Forschung gerückt sind. Es treten z.B. zeit- und gewebespezifische miRNA-Expressionsmuster bei der Entwicklung von Pflanzen und Tieren oder bei der Regulation von physiologischen Prozessen wie der Apoptose, der Zellteilung und der Zelldifferenzierung auf. Da häufig eine Vielzahl von miRNAs an der Regulation eines Prozesses beteiligt sind, ist es notwendig, die Expression der verschiedenen miRNAs gleichzeitig zu messen, um so die entsprechenden Expressionmuster zu finden. Hierzu eignen sich miRNA-Mikroarrays, deren Auswertung im Rahmen dieser Arbeit etabliert wurde. Als Basis hierfür wurden die in Kapitel 2 beschriebenen Methoden verwendet. Der hier beschriebene miRNA-Mikroarray wurde zur Untersuchung der N-Myc abhängigen miRNA Expression *in vivo* und *in vitro* eingesetzt. Die Ergebnisse beider Experimente zeigen eine deutliche Überlappung auf, somit konnte sowohl in den Vorversuchen, die im Rahmen der Etablierung durchgeführt wurden, als auch in den Experimenten gezeigt werden, dass sich die miRNA Mikroarray-Plattform für die experimentelle Anwendung eignet. Es wurde jedoch auch deutlich, dass durch eine Optimierung des derzeitigen

Array-Designs, andere Normalisierungsmethoden angewendet werden könnten, die zu einer besseren Reduzierung von systematischen Fehlern führen könnten.

Mit der RNAi-Screening Technologie steht eine weitere Hochdurchsatz-Technologie zur Verfügung, die eine Untersuchung der direkten Interaktion verschiedener Genprodukte ermöglicht. Die Screens generieren sehr große Datenmengen, die häufig eine sehr hohe Variabilität ausweisen, somit ebenfalls eine effiziente Datenanalyse erfordern. In dieser Arbeit wurde anhand der Auswertung eines shRNA-Screens, mit dem der Einfluss verschiedener Kinasen auf die Stabilität von c-Myc getestet werden sollte, die Etablierung einer Auswertungsroutine beschrieben.

8 Summary

High throughput methods have become an elementary tool in molecular biological research for the elucidation of complex cellular processes. An important aspect in using high throughput technologies is the implementation of appropriate algorithms for data analysis. The lack of standardised data validation and quality assurance processes has been recognised in the past as one of the major problems for successfully implementing high throughput experimental technologies (Kaul 2005). These processes are therefore of particular importance for the success of these experiments. For DNA-Chip technology, which has been used for studying gene expression since the beginning of the nineties, standardised methods for data analysis have already been established. An overview of these methods is given in chapter 2. However, despite efficient data analysis the interpretation of these data in the context of the biological question remains difficult. Therefore, during the last years the main focus of interest was more on bioinformatic methods regarding the functional interpretation of the data. In chapter 3 some of these methods are described and then applied to the results of a microarray experiment studying the c-Myc dependent gene expression in T-lymphocytes of transgenic mice. It was shown that the applied methods are suitable for the functional interpretation of gene lists. As different methods lead to different results and also focus on different aspects it is expedient to use them to maximise the efficiency of the interpretation. One aspect of the functional analysis of a set of co-regulated genes is their possible regulation through a common transcription factor. A method for the detection of significant over-represented cis-regulatory motifs in the promoter sequence of a set of co-regulated genes has been established within this work. This method includes the following steps:

- Establishing of a database containing orthologous promoter sequences (mouse/human)
- Masking of repetitive parts of the sequences
- Alignment of orthologous sequences
- Examination the conserved sequences for transcription factor binding sites (TFBS)

- Correction of the number of binding sites regarding the length of the conserved promoter sequences
- Testing a group of co-regulated genes for the enrichment of TFBS in the promoter sequences compared to a background set

This method was applied to the microarray dataset already used in chapter 3. Position weight matrices (PWM) with E-box motifs, which already have been described as a binding motif for c-Myc were identified as significantly overrepresented. Additionally, an enrichment of the PWM for the transcription factor YY1 can be detected, whose constitutive repression is reduced due to the over-expression of c-Myc (Austen, Cerni et al. 1998).

Beside the regulation of the gene expression, for example by transcription factors, there are a number of other factors that play an important role in the regulation of cellular processes. One of these factors are miRNAs, which over the last years have become more and more important in molecular biological research. miRNAs show e.g. time- and tissue-specific expression patterns in plant and animal development and are involved in the regulation of physiological processes such as apoptosis, cell division and cell differentiation. Since often a high number of miRNAs are involved in the regulation of one process, it is necessary to measure the expression of different miRNAs simultaneously to find the corresponding pattern. miRNA microarrays can be used to measure these expression patterns. Within this work the algorithms for the data analysis of a miRNA microarray platform are established based on the methods described in chapter 2. These algorithms were applied to two experiments that analyse the N-Myc dependent miRNA expression *in vivo* and *in vitro*. The results of both experiments have a clear overlap in miRNA expression patterns. The preliminary test as well as the experiments showed that the miRNA platform is suitable for the experimental use and that the data analysis routine leads to reasonable results. However it also became apparent that an improvement of the actual array design would allow different normalisation techniques which could lead to a better reduction of systematic errors.

With the RNAi screening technology another high throughput technology is available that allows the examination of the direct interaction of different gene products. These screens generate very large data sets which often show a high variability. This also necessitates efficient data analysis. In this work I describe the implementation of a data

analysis routine for the RNAi screening technology on the basis of an shRNA screen that examines the influence of different kinases on the stability of c-Myc.

9 Literaturverzeichnis

(2001). "Creating the gene ontology resource: design and implementation." Genome Res **11**(8): 1425-33.

(2006). "The Gene Ontology (GO) project in 2006." Nucleic Acids Res **34**(Database issue): D322-6.

Adhikary, S. and M. Eilers (2005). "Transcriptional regulation and transformation by Myc proteins." Nat Rev Mol Cell Biol **6**(8): 635-45.

Alland, L., R. Muhle, et al. (1997). "Role for N-CoR and histone deacetylase in Sin3-mediated transcriptional repression." **387**(6628): 49-55.

Alvarez, E., I. C. Northwood, et al. (1991). "Pro-Leu-Ser/Thr-Pro is a consensus primary sequence for substrate protein phosphorylation. Characterization of the phosphorylation of c-myc and c-jun proteins by an epidermal growth factor receptor threonine 669 protein kinase." J Biol Chem **266**(23): 15277-85.

Ambros, V. (2004). "The functions of animal microRNAs." Nature **431**(7006): 350-5.

Aparicio, S. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." **297**: 1301-1310.

Arabi, A., S. Wu, et al. (2005). "c-Myc associates with ribosomal DNA and activates RNA polymerase I transcription." Nat Cell Biol **7**(3): 303-10.

Aravin, A. A., M. Lagos-Quintana, et al. (2003). "The small RNA profile during Drosophila melanogaster development." Dev Cell **5**(2): 337-50.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Austen, M., C. Cerni, et al. (1998). "YY1 can inhibit c-Myc function through a mechanism requiring DNA binding of YY1 but neither its transactivation domain nor direct interaction with c-Myc." Oncogene **17**(4): 511-20.

Ayer, D. E., Q. A. Lawrence, et al. (1995). "Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3." Cell **80**(5): 767-76.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-97.

Baudino, T. A., C. McKay, et al. (2002). "c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression." Genes Dev **16**(19): 2530-43.

Bello-Fernandez, C., G. Packham, et al. (1993). "The ornithine decarboxylase gene is a transcriptional target of c-Myc." Proc Natl Acad Sci U S A **90**(16): 7804-8.

Bender, R., S. Lange, et al. (2007). "[Multiple testing]." Dtsch Med Wochenschr **132 Suppl 1**: e26-9.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B. Methodological **1**: 289-300.

Bernards, R., T. R. Brummelkamp, et al. (2006). "shRNA libraries and their use in cancer genetics." Nat Methods **3**(9): 701-6.

Blackwell, T. K., J. Huang, et al. (1993). "Binding of myc proteins to canonical and noncanonical DNA sequences." Mol Cell Biol **13**(9): 5216-24.

Boon, K., H. N. Caron, et al. (2001). "N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis." Embo J **20**(6): 1383-93.

Bouchard, C., O. Dittrich, et al. (2001). "Regulation of cyclin D2 gene expression by the Myc/Max/Mad network: Myc-dependent TRRAP recruitment and histone acetylation at the cyclin D2 promoter." Genes Dev **15**(16): 2042-7.

Bouchard, C., K. Thieke, et al. (1999). "Direct induction of cyclin D2 by Myc contributes to cell cycle progression and sequestration of p27." Embo J **18**(19): 5321-33.

Bowen, H., T. E. Biggs, et al. (2002). "c-Myc represses the murine Nramp1 promoter." Biochem Soc Trans **30**(4): 774-7.

Bravo, R., J. Burckhardt, et al. (1985). "Stimulation and inhibition of growth by EGF in different A431 cell clones is accompanied by the rapid induction of c-fos and c-myc proto-oncogenes." Embo J **4**(5): 1193-7.

Brideau, C., B. Gunter, et al. (2003). "Improved statistical methods for hit selection in high-throughput screening." J Biomol Screen **8**(6): 634-47.

Brodeur, G. M., R. C. Seeger, et al. (1984). "Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage." Science **224**(4653): 1121-4.

Brown, C. S., P. C. Goodwin, et al. (2001). "Image metrics in the statistical analysis of DNA microarray data." Proc Natl Acad Sci U S A **98**(16): 8944-9.

Brown, T. A. (2002). Genomes. New York and London, Garland Science.

Brummelkamp, T. R., R. Bernards, et al. (2002). "A system for stable expression of short interfering RNAs in mammalian cells." Science **296**(5567): 550-3.

Bucher, P. (1990). "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences." J Mol Biol **212**(4): 563-78.

Buchholz, M., A. Schatz, et al. (2006). "Overexpression of c-myc in pancreatic cancer caused by ectopic activation of NFATc1 and the Ca2+/calcineurin signaling pathway." Embo J **25**(15): 3714-24.

Calin, G. A., M. Ferracin, et al. (2005). "A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia." N Engl J Med **353**(17): 1793-801.

Caplen, N. J., S. Parrish, et al. (2001). "Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems." Proc Natl Acad Sci U S A **98**(17): 9742-7.

Carroll, J. S., C. A. Meyer, et al. (2006). "Genome-wide analysis of estrogen receptor binding sites." Nat Genet **38**(11): 1289-97.

Cawley, S., S. Bekiranov, et al. (2004). "Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs." Cell **116**(4): 499-509.

Chan, J. A., A. M. Krichevsky, et al. (2005). "MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells." Cancer Res **65**(14): 6029-33.

Chen, L. (1999). "Combinatorial gene regulation by eukaryotic transcription factors." Curr Opin Struct Biol **9**(1): 48-55.

Chen, Q. K., G. Z. Hertz, et al. (1995). "MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices." Comput Appl Biosci **11**(5): 563-6.

Chu, C. Y. and T. M. Rana (2007). "Small RNAs: regulators and guardians of the genome." J Cell Physiol **213**(2): 412-9.

Churchill, G. A. (2002). "Fundamentals of experimental design for cDNA microarrays." Nat Genet **32 Suppl**: 490-5.

Ciafre, S. A., S. Galardi, et al. (2005). "Extensive modulation of a set of microRNAs in primary glioblastoma." Biochem Biophys Res Commun **334**(4): 1351-8.

Coller, H. A., C. Grandori, et al. (2000). "Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion." Proc Natl Acad Sci U S A **97**(7): 3260-5.

Crabtree, G. R. and E. N. Olson (2002). "NFAT signaling: choreographing the social lives of cells." Cell **109 Suppl**: S67-79.

Croce, C. M. and G. A. Calin (2005). "miRNAs, cancer, and stem cell division." Cell **122**(1): 6-7.

Cummins, J. M. and V. E. Velculescu (2006). "Implications of micro-RNA profiling for cancer diagnosis." Oncogene **25**(46): 6220-7.

Dalla-Favera, R., E. P. Gelmann, et al. (1982). "Cloning and characterization of different human sequences related to the onc gene (v-myc) of avian myelocytomatosis virus (MC29)." Proc Natl Acad Sci U S A **79**(21): 6497-501.

Davuluri, R. V., I. Grosse, et al. (2001). "Computational identification of promoters and first exons in the human genome." Nat Genet **29**(4): 412-7.

Dennis, P. P. and A. Omer (2005). "Small non-coding RNAs in Archaea." Curr Opin Microbiol **8**(6): 685-94.

Dermitzakis, E. T. and A. G. Clark (2002). "Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover." **19**: 1114-1121.

Dietzl, G., D. Chen, et al. (2007). "A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila." Nature **448**(7150): 151-6.

Donjerkovic, D. and D. W. Scott (2000). "Regulation of the G1 phase of the mammalian cell cycle." Cell Res **10**(1): 1-16.

Eilers, M., D. Picard, et al. (1989). "Chimaeras of myc oncoprotein and steroid receptors cause hormone-dependent transformation of cells." Nature **340**(6228): 66-8.

Elbashir, S. M., J. Harborth, et al. (2001). "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells." Nature **411**(6836): 494-8.

Elnitski, L. (2003). "Distinguishing regulatory DNA from neutral sites." **13**: 64-72.

Euskirchen, G., T. E. Royce, et al. (2004). "CREB binds to multiple loci on human chromosome 22." Mol Cell Biol **24**(9): 3804-14.

Fernandez, P. C., S. R. Frank, et al. (2003). "Genomic targets of the human c-Myc protein." Genes Dev **17**(9): 1115-29.

Fire, A., S. Xu, et al. (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature **391**(6669): 806-11.

Frye, M., C. Gardner, et al. (2003). "Evidence that Myc activation depletes the epidermal stem cell compartment by modulating adhesive interactions with the local microenvironment." Development **130**(12): 2793-808.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-82.

Gomez-Roman, N., C. Grandori, et al. (2003). "Direct activation of RNA polymerase III transcription by c-Myc." Nature **421**(6920): 290-4.

Gottesman, S. (2004). "The small RNA regulators of Escherichia coli: roles and mechanisms*." Annu Rev Microbiol **58**: 303-28.

Grandori, C., N. Gomez-Roman, et al. (2005). "c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I." Nat Cell Biol **7**(3): 311-8.

Greasley, P. J., C. Bonnard, et al. (2000). "Myc induces the nucleolin and BN51 genes: possible implications in ribosome biogenesis." Nucleic Acids Res **28**(2): 446-53.

Gregory, R. I. and R. Shiekhattar (2005). "MicroRNA biogenesis and cancer." Cancer Res **65**(9): 3509-12.

Grewal, S. S., L. Li, et al. (2005). "Myc-dependent regulation of ribosomal RNA synthesis during Drosophila development." Nat Cell Biol **7**(3): 295-302.

Gubler, U. and B. J. Hoffman (1983). "A simple and very efficient method for generating cDNA libraries." Gene **25**(2-3): 263-9.

Gunter, B., C. Brideau, et al. (2003). "Statistical and graphical methods for quality control determination of high-throughput screening data." J Biomol Screen **8**(6): 624-33.

Guo, Q. M., R. L. Malek, et al. (2000). "Identification of c-myc responsive genes using rat cDNA microarray." Cancer Res **60**(21): 5922-8.

Hardison, R., J. L. Slightom, et al. (1997). "Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights." Gene **205**(1-2): 73-94.

Hartman, S. E., P. Bertone, et al. (2005). "Global changes in STAT target selection and transcription regulation upon interferon treatments." Genes Dev **19**(24): 2953-68.

Hartmann, O. (2005). "Quality control for microarray experiments." Methods Inf Med **44**(3): 408-13.

He, L. and G. J. Hannon (2004). "MicroRNAs: small RNAs with a big role in gene regulation." Nat Rev Genet **5**(7): 522-31.

Hermeking, H., C. Rago, et al. (2000). "Identification of CDK4 as a target of c-MYC." Proc Natl Acad Sci U S A **97**(5): 2229-34.

Heruth, D. P., L. A. Wetmore, et al. (1995). "Influence of protein tyrosine phosphorylation on the expression of the c-myc oncogene in cancer of the large bowel." J Cell Biochem **58**(1): 83-94.

Heuer C., H. T., Prause B. (2002). "A novel approach for quality control and correction of HTS data based on artificial intelligence." The Pharmaceutical Discovery & Development Report 2003/03, PharmaVentures Ltd **Retrieved from http://www.worldpharmaweb.com/pdd/new/overview5.pdf**.

Hochberg, Y. and Y. Benjamini (1990). "More powerful procedures for multiple significance testing." Stat Med **9**(7): 811-8.

Houbaviy, H. B., M. F. Murray, et al. (2003). "Embryonic stem cell-specific MicroRNAs." Dev Cell **5**(2): 351-8.

Ioshikhes, I. P. and M. Q. Zhang (2000). "Large-scale human promoter mapping using CpG islands." Nat Genet **26**(1): 61-3.

Iversen, P. W., B. J. Eastwood, et al. (2006). "A comparison of assay performance measures in screening assays: signal window, Z' factor, and assay variability ratio." J Biomol Screen **11**(3): 247-52.

Iyer, V. R., C. E. Horak, et al. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature **409**(6819): 533-8.

Jaattela, M., D. Wissing, et al. (1992). "Major heat shock protein hsp70 protects tumor cells from tumor necrosis factor cytotoxicity." Embo J **11**(10): 3507-12.

Jagadeesh, S., S. Kyo, et al. (2006). "Genistein represses telomerase activity via both transcriptional and posttranslational mechanisms in human prostate cancer cells." Cancer Res **66**(4): 2107-15.

Johnson, S. M., H. Grosshans, et al. (2005). "RAS is regulated by the let-7 microRNA family." Cell **120**(5): 635-47.

Jolly, E. R., C. S. Chin, et al. (2005). "Genome-wide identification of the regulatory targets of a transcription factor using biochemical characterization and computational genomic analysis." BMC Bioinformatics **6**: 275.

Kaneto, H., A. Sharma, et al. (2002). "Induction of c-Myc expression suppresses insulin gene transcription by inhibiting NeuroD/BETA2-mediated transcriptional activation." J Biol Chem **277**(15): 12998-3006.

Kaul, A. (2005). "The impact of sophisticated data analysis on the drug discovery process." Future Drug Discovery.

Kawano, M., A. A. Reynolds, et al. (2005). "Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli." Nucleic Acids Res **33**(3): 1040-50.

Kel, A. E., E. Gossling, et al. (2003). "MATCH: A tool for searching transcription factor binding sites in DNA sequences." Nucleic Acids Res **31**(13): 3576-9.

Kim, J. W., K. I. Zeller, et al. (2004). "Evaluation of myc E-box phylogenetic footprints in glycolytic genes by chromatin immunoprecipitation assays." Mol Cell Biol **24**(13): 5923-36.

Kim, S., Q. Li, et al. (2000). "Induction of ribosomal genes and hepatocyte hypertrophy by adenovirus-mediated expression of c-Myc in vivo." Proc Natl Acad Sci U S A **97**(21): 11198-202.

Kimura, K., A. Wakamatsu, et al. (2006). "Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes." Genome Res **16**(1): 55-65.

Knudsen, S. (1999). "Promoter2.0: for the recognition of PolII promoter sequences." Bioinformatics **15**(5): 356-61.

Kumagai, T., N. Wakimoto, et al. (2007). "Histone deacetylase inhibitor, suberoylanilide hydroxamic acid (Vorinostat, SAHA) profoundly inhibits the growth of human pancreatic cancer cells." Int J Cancer **121**(3): 656-65.

Lagos-Quintana, M., R. Rauhut, et al. (2001). "Identification of novel genes coding for small expressed RNAs." Science **294**(5543): 853-8.

Lagos-Quintana, M., R. Rauhut, et al. (2002). "Identification of tissue-specific microRNAs from mouse." Curr Biol **12**(9): 735-9.

Lai, E. C., P. Tomancak, et al. (2003). "Computational identification of Drosophila microRNA genes." Genome Biol **4**(7): R42.

Lamb, J. (2007). "The Connectivity Map: a new tool for biomedical research." Nat Rev Cancer **7**(1): 54-60.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lau, N. C., L. P. Lim, et al. (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-62.

Lee, R. C. and V. Ambros (2001). "An extensive class of small RNAs in Caenorhabditis elegans." Science **294**(5543): 862-4.

Lee, R. C., R. L. Feinbaum, et al. (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell **75**(5): 843-54.

Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." Embo J **23**(20): 4051-60.

Lenhard, B., A. Sandelin, et al. (2003). "Identification of conserved regulatory elements by comparative genome analysis." J Biol **2**(2): 13.

Lewis, B. P., C. B. Burge, et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Li, H. and X. Wu (2004). "Histone deacetylase inhibitor, Trichostatin A, activates p21WAF1/CIP1 expression through downregulation of c-myc and release of the repression of c-myc from the promoter in human cervical cancer cells." Biochem Biophys Res Commun **324**(2): 860-7.

Li, Z., S. Van Calcar, et al. (2003). "A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells." Proc Natl Acad Sci U S A **100**(14): 8164-9.

Lieb, J. D., X. Liu, et al. (2001). "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association." Nat Genet **28**(4): 327-34.

Lim, L. P., N. C. Lau, et al. (2003). "The microRNAs of Caenorhabditis elegans." Genes Dev **17**(8): 991-1008.

Littlewood, T. D., D. C. Hancock, et al. (1995). "A modified oestrogen receptor ligand-binding domain as an improved switch for the regulation of heterologous proteins." Nucleic Acids Res **23**(10): 1686-90.

Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.

Loots, G. G., R. M. Locksley, et al. (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons." Science **288**(5463): 136-40.

Loots, G. G., I. Ovcharenko, et al. (2002). "rVista for comparative sequence-based discovery of functional transcription factor binding sites." Genome Res **12**(5): 832-9.

Lu, J., J. Qian, et al. (2005). "Differential expression of components of the microRNA machinery during mouse organogenesis." Biochem Biophys Res Commun **334**(2): 319-23.

Mao, D. Y. L., J. D. Watson, et al. (2003). "Analysis of Myc Bound Loci Identified by CpG Island Arrays Shows that Max Is Essential for Myc-Dependent Repression." Current Biology **13**(10): 882-886.

Marcotte, E. M., M. Pellegrini, et al. (1999). "A combined algorithm for genome-wide prediction of protein function." Nature **402**(6757): 83-6.

Martone, R., G. Euskirchen, et al. (2003). "Distribution of NF-kappaB-binding sites across human chromosome 22." Proc Natl Acad Sci U S A **100**(21): 12247-52.

Mateyak, M. K., A. J. Obaya, et al. (1997). "Phenotypes of c-Myc-deficient rat fibroblasts isolated by targeted homologous recombination." Cell Growth Differ **8**(10): 1039-48.

Mattick, J. S. (2003). "Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms." Bioessays **25**(10): 930-9.

Mattick, J. S. (2004). "RNA regulation: a new genetics?" Nat Rev Genet **5**(4): 316-23.

Mattick, J. S. and I. V. Makunin (2006). "Non-coding RNA." Hum Mol Genet **15 Spec No 1**: R17-29.

Menssen, A. and H. Hermeking (2002). "Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes." Proc Natl Acad Sci U S A **99**(9): 6274-9.

Metzler, M., M. Wilda, et al. (2004). "High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma." Genes Chromosomes Cancer **39**(2): 167-9.

Michael, M. Z., O. C. SM, et al. (2003). "Reduced accumulation of specific microRNAs in colorectal neoplasia." Mol Cancer Res **1**(12): 882-91.

Miltenberger, R. J., K. A. Sukow, et al. (1995). "An E-box-mediated increase in cad transcription at the G1/S-phase boundary is suppressed by inhibitory c-Myc mutants." Mol Cell Biol **15**(5): 2527-35.

Miyagishi, M. and K. Taira (2002). "U6 promoter-driven siRNAs with four uridine 3' overhangs efficiently suppress targeted gene expression in mammalian cells." Nat Biotechnol **20**(5): 497-500.

Morgenstern, B. (1999). "DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment." Bioinformatics **15**(3): 211-8.

Morrish, F., C. Giedt, et al. (2003). "c-MYC apoptotic function is mediated by NRF-1 target genes." Genes Dev **17**(2): 240-55.

Neph, S. and M. Tompa (2006). "MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes." Nucleic Acids Res **34**(Web Server issue): W366-8.

Noma, K., T. Sugiyama, et al. (2004). "RITS acts in cis to promote RNA interference-mediated transcriptional and post-transcriptional silencing." Nat Genet **36**(11): 1174-80.

O'Connell, B. C., A. F. Cheung, et al. (2003). "A large scale genetic analysis of c-Myc-regulated gene expression patterns." J Biol Chem **278**(14): 12563-73.

O'Donnell, K. A., E. A. Wentzel, et al. (2005). "c-Myc-regulated microRNAs modulate E2F1 expression." Nature **435**(7043): 839-43.

Orian, A., B. van Steensel, et al. (2003). "Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network." Genes Dev **17**(9): 1101-14.

Oster, S. K., C. S. Ho, et al. (2002). "The myc oncogene: MarvelouslY Complex." Adv Cancer Res **84**: 81-154.

Osthus, R. C., H. Shim, et al. (2000). "Deregulation of glucose transporter 1 and glycolytic gene expression by c-Myc." J Biol Chem **275**(29): 21797-800.

Paddison, P. J., A. A. Caudy, et al. (2002). "Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells." Genes Dev **16**(8): 948-58.

Pahlman, S., L. Odelstad, et al. (1981). "Phenotypic changes of human neuroblastoma cells in culture induced by 12-O-tetradecanoyl-phorbol-13-acetate." Int J Cancer **28**(5): 583-9.

Pasquinelli, A. E., B. J. Reinhart, et al. (2000). "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA." Nature **408**(6808): 86-9.

Paul, C. P., P. D. Good, et al. (2002). "Effective expression of small interfering RNA in human cells." Nat Biotechnol **20**(5): 505-8.

Pollard, D. A., C. M. Bergman, et al. (2004). "Benchmarking tools for the alignment of functional noncoding DNA." BMC Bioinformatics **5**: 6.

Poortinga, G., K. M. Hannan, et al. (2004). "MAD1 and c-MYC regulate UBF and rDNA transcription during granulocyte differentiation." Embo J **23**(16): 3325-35.

Prestridge, D. S. (1995). "Predicting Pol II promoter sequences using transcription factor binding sites." J Mol Biol **249**(5): 923-32.

Prestridge, D. S. (1996). "SIGNAL SCAN 4.0: additional databases and sequence formats." Comput Appl Biosci **12**(2): 157-60.

Qin, L., P. Qiu, et al. (2003). "Gene expression profiles and transcription factors involved in parathyroid hormone signaling in osteoblasts revealed by microarray and bioinformatics." J Biol Chem **278**(22): 19723-31.

Qiu, P., W. Ding, et al. (2002). "Computational analysis of composite regulatory elements." Mamm Genome **13**(6): 327-32.

Quandt, K., K. Frech, et al. (1995). "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data." Nucleic Acids Res **23**(23): 4878-84.

Ramaswamy, S., P. Tamayo, et al. (2001). "Multiclass cancer diagnosis using tumor gene expression signatures." Proc Natl Acad Sci U S A **98**(26): 15149-54.

Ravasi, T., H. Suzuki, et al. (2006). "Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome." Genome Res **16**(1): 11-9.

Ren, B., F. Robert, et al. (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-9.

Riggs, K. J., S. Saleque, et al. (1993). "Yin-yang 1 activates the c-myc promoter." Mol Cell Biol **13**(12): 7487-95.

Robanus-Maandag, E. C., C. A. Bosch, et al. (2003). "Association of C-MYC amplification with progression from the in situ to the invasive stage in C-MYC-amplified breast carcinomas." J Pathol **201**(1): 75-82.

Roulet, E., I. Fisch, et al. (1998). "Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA." In Silico Biol **1**(1): 21-8.

Roy, A. L., M. Meisterernst, et al. (1991). "Cooperative interaction of an initiator-binding transcription initiation factor and the helix-loop-helix activator USF." Nature **354**(6350): 245-8.

Sandelin, A., P. Carninci, et al. (2007). "Mammalian RNA polymerase II core promoters: insights from genome-wide studies." Nat Rev Genet **8**(6): 424-36.

Sankoff, D. and R. J. Cedergren (1973). "A test for nucleotide sequence homology." J Mol Biol **77**(1): 169-64.

Sauer, T., E. Shelest, et al. (2006). "Evaluating phylogenetic footprinting for human-rodent comparisons." Bioinformatics **22**(4): 430-7.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

Scherf, M., A. Klingenhoff, et al. (2000). "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach." J Mol Biol **297**(3): 599-606.

Schlosser, I., M. Holzel, et al. (2003). "A role for c-Myc in the regulation of ribosomal RNA processing." Nucleic Acids Res **31**(21): 6148-56.

Schug, J., W. P. Schuller, et al. (2005). "Promoter features related to tissue specificity as measured by Shannon entropy." Genome Biol **6**(4): R33.

Schuhmacher, M., F. Kohlhuber, et al. (2001). "The transcriptional program of a human B cell line in response to Myc." Nucleic Acids Res **29**(2): 397-406.

Schulte JH, Horn S, Otto T, Samans B, Heukamp LC, Eilers UC, Krause M, Astrahantseff K, Klein-Hitpass L, Buettner R, Schramm A, Christiansen H, Eilers M, Eggert A, Berwanger B. MYCN regulates oncogenic MicroRNAs in neuroblastoma. Int J Cancer. 2008 Feb 1;122(3):699-704.

Schwab, M., K. Alitalo, et al. (1983). "Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour." Nature **305**(5931): 245-8.

Schwartzberg, P. L., L. D. Finkelstein, et al. (2005). "TEC-family kinases: regulators of T-helper-cell differentiation." Nat Rev Immunol **5**(4): 284-95.

Seoane, J., C. Pouponnot, et al. (2001). "TGFbeta influences Myc, Miz-1 and Smad to control the CDK inhibitor p15INK4b." Nat Cell Biol **3**(4): 400-8.

Sheiness, D., L. Fanshier, et al. (1978). "Identification of nucleotide sequences which may encode the oncogenic capacity of avian retrovirus MC29." J Virol **28**(2): 600-10.

Sherr, C. J. and J. M. Roberts (1999). "CDK inhibitors: positive and negative regulators of G1-phase progression." Genes Dev **13**(12): 1501-12.

Shiio, Y., S. Donohoe, et al. (2002). "Quantitative proteomic analysis of Myc oncoprotein function." Embo J **21**(19): 5088-96.

Shingara, J., K. Keiger, et al. (2005). "An optimized isolation and labeling platform for accurate microRNA expression profiling." Rna **11**(9): 1461-70.

Shrivastava, A., J. Yu, et al. (1996). "YY1 and c-Myc associate in vivo in a manner that depends on c-Myc levels." Proc Natl Acad Sci U S A **93**(20): 10638-41.

Simon, R. M. and K. Dobbin (2003). "Experimental design of DNA microarray experiments." Biotechniques **Suppl**: 16-21.

Smale, S. T. and J. T. Kadonaga (2003). "The RNA polymerase II core promoter." Annu Rev Biochem **72**: 449-79.

Solovyev, V. and A. Salamov (1997). "The Gene-Finder computer tools for analysis of human and model organisms genome sequences." Proc Int Conf Intell Syst Mol Biol **5**: 294-302.

Sontheimer, E. J. and R. W. Carthew (2004). "MOLECULAR BIOLOGY: Argonaute Journeys into the Heart of RISC 10.1126/science.1103076." Science 305(5689): 1409-1410.

Steinbrink, S., H. T., et al. (2004). "RNA-Interferenz im Hochdurchsatz." BIOSpektrum 13. Jahrgang (Sonderheft zur Biotechnica 2007): 782-785.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Stormo, G. D., T. D. Schneider, et al. (1982). "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli." Nucleic Acids Res **10**(9): 2997-3011.

Storz, G., S. Altuvia, et al. (2005). "An abundance of RNA regulators." Annu Rev Biochem **74**: 199-217.

Storz, G., J. A. Opdyke, et al. (2004). "Controlling mRNA stability and translation with small, noncoding RNAs." Curr Opin Microbiol **7**(2): 140-4.

Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proc Natl Acad Sci U S A **101**(16): 6062-7.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-50.

Sui, G., C. Soohoo, et al. (2002). "A DNA vector-based RNAi technology to suppress gene expression in mammalian cells." Proc Natl Acad Sci U S A **99**(8): 5515-20.

Suzuki, Y., R. Yamashita, et al. (2004). "Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions." Genome Res **14**(9): 1711-8.

Tagle, D. A., B. F. Koop, et al. (1988). "Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints." J Mol Biol **203**(2): 439-55.

Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol **23**(1): 137-44.

Trinklein, N. D., S. J. Aldred, et al. (2003). "Identification and functional analysis of human transcriptional promoters." Genome Res **13**(2): 308-12.

Truss, M. and C. Hagemeier (2004). "Mechanismen der RNA-Interferenz in Säugerzellen." BIOSpektrum **4**: 782-785.

Tseng, G. C., M. K. Oh, et al. (2001). "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects." Nucleic Acids Res **29**(12): 2549-57.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.

Vennstrom, B., D. Sheiness, et al. (1982). "Isolation and characterization of c-myc, a cellular homolog of the oncogene (v-myc) of avian myelocytomatosis virus strain 29." J Virol **42**(3): 773-9.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Vogel, J., V. Bartels, et al. (2003). "RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria." Nucleic Acids Res **31**(22): 6435-43.

Volinia, S., G. A. Calin, et al. (2006). "A microRNA expression signature of human solid tumors defines cancer gene targets." Proc Natl Acad Sci U S A **103**(7): 2257-61.

Wang, X., S. Ghosh, et al. (2001). "Quantitative quality control in microarray image processing and data acquisition." Nucleic Acids Res **29**(15): E75-5.

Wang, Y., J. G. Klijn, et al. (2005). "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." Lancet **365**(9460): 671-9.

Wanzel, M., S. Herold, et al. (2003). "Transcriptional repression by Myc." Trends Cell Biol **13**(3): 146-50.

Wasserman, W. W. and J. W. Fickett (1998). "Identification of regulatory regions which confer muscle-specific gene expression." J Mol Biol **278**(1): 167-81.

Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Watson, J. D., S. K. Oster, et al. (2002). "Identifying genes regulated in a Myc-dependent manner." J Biol Chem **277**(40): 36921-30.

Wei, C. L., Q. Wu, et al. (2006). "A global map of p53 transcription-factor binding sites in the human genome." Cell **124**(1): 207-19.

Werner, T., S. Fessele, et al. (2003). "Computer modeling of promoter organization as a tool to study transcriptional coregulation." Faseb J **17**(10): 1228-37.

West, M. J., M. Stoneley, et al. (1998). "Translational induction of the c-myc oncogene via activation of the FRAP/TOR signalling pathway." Oncogene **17**(6): 769-80.

Wilderman, P. J., N. A. Sowa, et al. (2004). "Identification of tandem duplicate regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis." Proc Natl Acad Sci U S A **101**(26): 9792-7.

Wilson, A., M. J. Murphy, et al. (2004). "c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation." Genes Dev **18**(22): 2747-63.

Wonsey, D. R., K. I. Zeller, et al. (2002). "The c-Myc target gene PRDX3 is required for mitochondrial homeostasis and neoplastic transformation." Proc Natl Acad Sci U S A **99**(10): 6649-54.

Wu, K. J., A. Polack, et al. (1999). "Coordinated regulation of iron-controlling genes, H-ferritin and IRP2, by c-MYC." Science **283**(5402): 676-9.

Wu, S., C. Cetinkaya, et al. (2003). "Myc represses differentiation-induced p21CIP1 expression via Miz-1-dependent interaction with the p21 core promoter." Oncogene **22**(3): 351-60.

Xuan, Z., F. Zhao, et al. (2005). "Genome-wide promoter extraction and analysis in human, mouse, and rat." Genome Biol **6**(8): R72.

Xue, W., J. Wang, et al. (2004). "Enrichment of transcriptional regulatory sites in non-coding genomic region." Bioinformatics **20**(4): 569-75.

Yamashita, R., Y. Suzuki, et al. (2006). "DBTSS: DataBase of Human Transcription Start Sites, progress report 2006." Nucleic Acids Res **34**(Database issue): D86-9.

Yang, B. S., T. J. Geddes, et al. (1991). "Transcriptional suppression of cellular gene expression by c-Myc." Mol Cell Biol **11**(4): 2291-5.

Yang, B. S., J. D. Gilbert, et al. (1993). "Overexpression of Myc suppresses CCAAT transcription factor/nuclear factor 1-dependent promoters in vivo." Mol Cell Biol **13**(5): 3093-102.

Yang, Y. H., S. Dudoit, et al. (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." Nucleic Acids Res **30**(4): e15.

Yang, Y. H. and T. Speed (2002). "Design issues for cDNA microarray experiments." Nat Rev Genet **3**(8): 579-88.

Yeh, C. W., S. S. Huang, et al. (2006). "Ras-dependent recruitment of c-Myc for transcriptional activation of nucleophosmin/B23 in highly malignant U1 bladder cancer cells." Mol Pharmacol **70**(4): 1443-53.

Yu, J. Y., S. L. DeRuiter, et al. (2002). "RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells." Proc Natl Acad Sci U S A **99**(9): 6047-52.

Zeller, K. I., A. G. Jegga, et al. (2003). "An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets." Genome Biol **4**(10): R69.

Zhang, J. H., T. D. Chung, et al. (1999). "A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays." J Biomol Screen **4**(2): 67-73.

Zhang, M. Q. (1998). "Identification of human gene core promoters in silico." Genome Res **8**(3): 319-26.

Zhu, Z., Y. Pilpel, et al. (2002). "Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm." J Mol Biol **318**(1): 71-81.

10 Anhang

10.1 EASE-Auswertung — Liste der signifikant angereicherten funktionellen Themenbereiche infolge Aktivierung bzw. Reprimierung durch c-Myc

System	Gene Category	List Hits	List Total	Population Hits	Population Total	EASE score
KEGG pathway	MMU03010:RIBOSOME	44	193	66	2076	2.07E-30
GO Cellular Component	RIBOSOME	53	505	127	6591	5.30E-26
GO Molecular Function	STRUCTURAL CONSTITUENT OF RIBOSOME	53	550	130	7331	9.03E-26
GO Biological Process	BIOSYNTHESIS	125	495	701	6603	7.95E-22
GO Biological Process	CELLULAR BIOSYNTHESIS	112	495	615	6603	4.05E-20
GO Cellular Component	RIBONUCLEOPROTEIN COMPLEX	65	505	261	6591	6.33E-18
GO Biological Process	MACROMOLECULE BIOSYNTHESIS	83	495	416	6603	3.81E-17
GO Biological Process	PROTEIN BIOSYNTHESIS	78	495	376	6603	4.18E-17
GO Biological Process	RIBOSOME BIOGENESIS AND ASSEMBLY	31	495	77	6603	5.46E-15
GO Biological Process	CYTOPLASM ORGANIZATION AND BIOGENESIS	32	495	87	6603	3.48E-14
GO Cellular Component	CYTOSOLIC RIBOSOME (SENSU EUKARYOTA)	20	505	33	6591	1.32E-13
GO Molecular Function	STRUCTURAL MOLECULE ACTIVITY	61	550	296	7331	3.72E-13
GO Biological Process	RIBOSOME BIOGENESIS	27	495	69	6603	9.42E-13
GenMAPP pathway	Mm_Ribosomal Proteins	21	76	35	573	1.35E-10
KEGG pathway	HSA03010:RIBOSOME	12	193	16	2076	9.33E-09
GO Cellular Component	INTRACELLULAR NON-MEMBRANE-BOUND ORGANELLE	104	505	794	6591	1.47E-08
GO Cellular Component	NON-MEMBRANE-BOUND ORGANELLE	104	505	794	6591	1.47E-08
GO Cellular Component	NUCLEOLUS	23	505	83	6591	1.46E-07
GO Cellular Component	CYTOPLASM	208	505	2174	6591	4.61E-05
GO Cellular Component	CYTOSOL	36	505	236	6591	9.65E-05
GO Cellular Component	PROTEIN COMPLEX	117	505	1116	6591	0.0001589
GO Molecular Function	RNA BINDING	47	550	357	7331	0.0001688
GO Biological Process	RRNA PROCESSING	10	495	30	6603	0.0002352
GO Cellular Component	INTRACELLULAR	376	505	4451	6591	0.0002921

GO Biological Process	RRNA METABOLISM	10	495	31	6603	0.0003098
GO Molecular Function	NUCLEOTIDYLTRANSFERASE ACTIVITY	16	550	77	7331	0.0004903
GO Biological Process	HETEROCYCLE METABOLISM	10	495	33	6603	0.0005179
GO Molecular Function	DNA-DIRECTED RNA POLYMERASE II ACTIVITY	7	550	16	7331	0.0007172
GO Molecular Function	DNA-DIRECTED RNA POLYMERASE I ACTIVITY	7	550	16	7331	0.0007172
GO Molecular Function	DNA-DIRECTED RNA POLYMERASE III ACTIVITY	7	550	16	7331	0.0007172
GO Cellular Component	SMALL RIBOSOMAL SUBUNIT	8	505	22	6591	0.0009071
GO Biological Process	CELLULAR MACROMOLECULE METABOLISM	156	495	1674	6603	0.0010104
GO Biological Process	TRANSLATIONAL ELONGATION	7	495	18	6603	0.0014506
GO Cellular Component	LARGE RIBOSOMAL SUBUNIT	8	505	24	6591	0.0016096
GO Biological Process	T CELL DIFFERENTIATION	7	495	19	6603	0.001988
GO Biological Process	CELLULAR PROTEIN METABOLISM	152	495	1655	6603	0.0023258
GO Molecular Function	TRANSLATION FACTOR ACTIVITY, NUCLEIC ACID BINDING	17	550	99	7331	0.0026028
GO Cellular Component	CELL	467	505	5851	6591	0.0027951
GO Molecular Function	TRANSLATION REGULATOR ACTIVITY	17	550	100	7331	0.0028903
GO Cellular Component	INTRACELLULAR ORGANELLE	322	505	3809	6591	0.002891
GO Cellular Component	ORGANELLE	322	505	3815	6591	0.0032922
GO Biological Process	TRANSLATION	18	495	112	6603	0.0037535
GO Molecular Function	DNA-DIRECTED RNA POLYMERASE ACTIVITY	8	550	28	7331	0.0037584
KEGG pathway	MMU00240:PYRIMIDINE METABOLISM	14	193	63	2076	0.0039404
GO Biological Process	MACROMOLECULE METABOLISM	207	495	2384	6603	0.0043556
GO Molecular Function	TRANSFERASE ACTIVITY, TRANSFERRING ONE-CARBON GROUPS	16	550	95	7331	0.0043908
GO Biological Process	REGULATION OF T CELL DIFFERENTIATION	5	495	10	6603	0.0045159
GO Biological Process	LYMPHOCYTE DIFFERENTIATION	9	495	37	6603	0.0051618
GO Cellular Component	NUCLEAR LUMEN	41	505	349	6591	0.0058251
Chromosome	Mus musculus 10	45	644	428	9168	0.0063581
GO Molecular Function	RRNA BINDING	5	550	11	7331	0.0067102
GO Biological Process	T CELL ACTIVATION	10	495	47	6603	0.0071671
KEGG pathway	MMU03020:RNA POLYMERASE	7	193	20	2076	0.0074333
KEGG pathway	MMU04612:ANTIGEN PROCESSING AND PRESENTATION	8	193	26	2076	0.0074367
GO Cellular Component	CYTOSOLIC LARGE RIBOSOMAL SUBUNIT (SENSU EUKARYOTA)	4	505	6	6591	0.0074623

GO Biological Process	PROTEIN METABOLISM	155	495	1742	6603	0.0076096
GenMAPP pathway	Mm_Translation Factors	9	76	24	573	0.0076179
GO Cellular Component	MEMBRANE-ENCLOSED LUMEN	46	505	411	6591	0.0081792
GO Cellular Component	ORGANELLE LUMEN	46	505	411	6591	0.0081792
GO Molecular Function	METHYLTRANSFERASE ACTIVITY	15	550	93	7331	0.0088921
GO Biological Process	TRNA PROCESSING	7	495	26	6603	0.0107808
GO Biological Process	CELLULAR METABOLISM	307	495	3763	6603	0.011477
KEGG pathway	MMU00230:PURINE METABOLISM	17	193	95	2076	0.0115746
GO Biological Process	NUCLEOBASE METABOLISM	4	495	7	6603	0.0115995
GO Biological Process	REGULATION OF LYMPHOCYTE DIFFERENTIATION	5	495	13	6603	0.0128395
Chromosome	Mus musculus Y	3	644	3	9168	0.0140485
GO Biological Process	PURINE BASE METABOLISM	3	495	3	6603	0.0159274
GO Molecular Function	NUCLEASE ACTIVITY	14	550	90	7331	0.0159501
GO Molecular Function	GLYCINE HYDROXYMETHYLTRANSFERASE ACTIVITY	3	550	3	7331	0.0159603
GO Biological Process	POSITIVE REGULATION OF T CELL DIFFERENTIATION	4	495	8	6603	0.0175398
GO Biological Process	HEMOPOIETIC OR LYMPHOID ORGAN DEVELOPMENT	14	495	92	6603	0.0187012
GO Biological Process	TRNA METABOLISM	11	495	64	6603	0.0194788
GO Molecular Function	TRANSLATION INITIATION FACTOR ACTIVITY	11	550	64	7331	0.0196529
GO Molecular Function	RIBONUCLEASE ACTIVITY	7	550	30	7331	0.0217019
GO Molecular Function	3'-5' EXONUCLEASE ACTIVITY	5	550	15	7331	0.0218406
GO Biological Process	LYMPHOCYTE ACTIVATION	12	495	75	6603	0.0227618
GO Biological Process	ORGANELLE ORGANIZATION AND BIOGENESIS	56	495	565	6603	0.02327
GO Biological Process	CELL ACTIVATION	13	495	85	6603	0.0234954
GO Biological Process	IMMUNE CELL ACTIVATION	13	495	85	6603	0.0234954
GO Biological Process	POSITIVE REGULATION OF LYMPHOCYTE DIFFERENTIATION	4	495	9	6603	0.0248705
GO Biological Process	PRIMARY METABOLISM	295	495	3645	6603	0.024917
GO Biological Process	IMMUNE RESPONSE	25	495	211	6603	0.0249828
KEGG pathway	MMU00440:AMINOPHOSPHONATE METABOLISM	5	193	13	2076	0.026022
GO Cellular Component	EUKARYOTIC 48S INITIATION COMPLEX	4	505	9	6591	0.0263553
GO Cellular Component	CYTOSOLIC SMALL RIBOSOMAL SUBUNIT (SENSU EUKARYOTA)	4	505	9	6591	0.0263553
GO Biological Process	PIGMENT BIOSYNTHESIS	5	495	16	6603	0.0273353

GO Biological Process	RNA PROCESSING	25	495	213	6603	0.0276109
GO Biological Process	NUCLEOTIDE BIOSYNTHESIS	11	495	68	6603	0.028668
GO Biological Process	METABOLISM	320	495	3995	6603	0.0297124
GO Molecular Function	EXONUCLEASE ACTIVITY	8	550	41	7331	0.0307574
Chromosome	Mus musculus X	31	644	302	9168	0.0329637
GO Biological Process	RNA METABOLISM	30	495	273	6603	0.0335246
GO Biological Process	CELL PROLIFERATION	24	495	206	6603	0.0335691
GO Molecular Function	S-ADENOSYLMETHIONINE-DEPENDENT METHYLTRANSFERASE ACTIVITY	9	550	51	7331	0.0343119
KEGG pathway	MMU00626:NITROBENZENE DEGRADATION	4	193	9	2076	0.0430488
KEGG pathway	MMU00670:ONE CARBON POOL BY FOLATE	4	193	9	2076	0.0430488
GO Molecular Function	UNFOLDED PROTEIN BINDING	13	550	93	7331	0.0438756
GO Biological Process	AROMATIC COMPOUND METABOLISM	9	495	54	6603	0.0458722
GO Cellular Component	NUCLEUS	188	505	2215	6591	0.047201
GO Biological Process	RIBOSOMAL PROTEIN IMPORT INTO NUCLEUS	3	495	5	6603	0.0479947
GO Biological Process	PIGMENT METABOLISM	5	495	19	6603	0.0487716

10.2 GSEA Auswertung – Liste der Gensets, die eine signifikante Anreicherung von Genen aufweist, die durch c-Myc hochreguliert werden

Rang	**NAME**	**SIZE**	**ES**	**NES**	**NOM p-val**	**FDR q-val**
1	GNF2_FBL	102	-0.7001	-2.2207	0	0
2	MORF_TPT1	69	-0.7289	-2.1744	0	0
3	MORF_ACTG1	96	-0.6402	-2.0798	0	0.00046
4	GNF2_ST13	48	-0.796	-2.0884	0	0.00053
5	MORF_NPM1	113	-0.6553	-2.1208	0	0.00061
6	GNF2_EIF3S6	84	-0.77	-2.0999	0	0.00064
7	GCM_NPM1	83	-0.6957	-2.1096	0	0.0008
8	FLOTHO_CASP8AP2_MRD_DIFF	40	-0.5554	-2.0488	0	0.00112
9	MORF_JUND	47	-0.6059	-2.0437	0	0.00172
10	GNF2_NPM1	44	-0.6949	-2.0343	0	0.00188
11	GNF2_DAP3	88	-0.6059	-2.0262	0	0.00189
12	MORF_NME2	108	-0.6034	-2.0138	0	0.00199
13	RIBOSOMAL_PROTEINS	60	-0.7807	-2.0095	0	0.00207
14	GCM_TPT1	45	-0.8338	-1.9944	0	0.00257
15	V$MYCMAX_01	133	-0.5495	-1.9995	0	0.00275

16	GNF2_RBBP6	47	-0.5614	-1.9744	0	0.00326
17	V$USF_C	140	-0.4808	-1.9752	0	0.00345
18	ZELLER_MYC_UP	18	-0.7532	-1.982	0	0.00348
19	MORF_DEAF1	41	-0.6887	-1.9623	0	0.00433
20	V$NMYC_01	133	-0.4631	-1.9493	0	0.00494
21	MORF_FBL	103	-0.601	-1.9402	0	0.00586
22	GNF2_GLTSCR2	25	-0.8555	-1.9311	0	0.00655
23	MYC_TARGETS	32	-0.7545	-1.9271	0	0.00691
24	AGUIRRE_PANCREAS_CHR17	30	-0.5466	-1.916	0	0.00809
25	SHEPARD_CRASH_AND_BURN_MUT_VS_WT_UP	90	-0.4857	-1.8976	0	0.00948
26	HESS_HOXAANMEIS1_DN	49	-0.5975	-1.8923	0.0039	0.00976
27	HESS_HOXAANMEIS1_UP	49	-0.5975	-1.8923	0.0039	0.01012
28	GCM_PSME1	68	-0.5532	-1.8935	0	0.01014
29	SCHUMACHER_MYC_UP	34	-0.6739	-1.8776	0	0.01137
30	LEE_MYC_UP	37	-0.5301	-1.8698	0	0.01264
31	MORF_EIF3S6	90	-0.5378	-1.8636	0	0.01284
32	TRANSLATION_FACTORS	39	-0.6632	-1.8655	0	0.01289
33	JAIN_NEMO_DIFF	45	-0.5556	-1.8618	0	0.01293
34	GCM_APEX1	88	-0.5586	-1.8659	0	0.01314
35	PENG_RAPAMYCIN_DN	157	-0.4986	-1.8527	0.0041	0.01425
36	HSC_EARLYPROGENITORS_ADULT	230	-0.4268	-1.8408	0	0.01558
37	UVB_NHEK1_C1	27	-0.6422	-1.8359	0.0039	0.01581
38	HSC_INTERMEDIATEPROGENITORS_ADULT	70	-0.528	-1.8378	0	0.01585
39	HSC_EARLYPROGENITORS_SHARED	228	-0.4257	-1.8413	0	0.01591
40	GNF2_MBD4	16	-0.7373	-1.8339	0	0.01597
41	HSC_EARLYPROGENITORS_FETAL	228	-0.4257	-1.8413	0	0.01635
42	HSC_INTERMEDIATEPROGENITORS_FETAL	74	-0.5434	-1.8245	0	0.01713
43	SANSOM_APC_LOSS5_UP	60	-0.4947	-1.8229	0	0.01715
44	CANTHARIDIN_DN	40	-0.6402	-1.8263	0	0.01729
45	GNF2_TPT1	26	-0.8201	-1.8274	0	0.01738
46	MORF_EIF4A2	89	-0.4977	-1.8159	0	0.0186
47	GNF2_UBE2I	31	-0.6018	-1.8141	0	0.01868
48	HSC_INTERMEDIATEPROGENITORS_SHARED	62	-0.5482	-1.7969	0	0.02207
49	STEMCELL_COMMON_UP	120	-0.4412	-1.7976	0	0.02232
50	MORF_UBE2I	166	-0.5034	-1.7983	0	0.02243
51	GCM_ACTG1	98	-0.5186	-1.7908	0	0.02289
52	MORF_RFC4	118	-0.4959	-1.7829	0.0078	0.02471
53	MORF_UNG	59	-0.5582	-1.7837	0.0075	0.02486
54	MORF_DNMT1	91	-0.5102	-1.7767	0.0075	0.02515
55	MENSSEN_MYC_UP	25	-0.6755	-1.7779	0	0.02537
56	CHR12Q23	20	-0.5948	-1.7799	0.0168	0.02543
57	AACWWCAANK_UNKNOWN	66	-0.4451	-1.7788	0	0.02554
58	V$MYC_Q2	87	-0.438	-1.768	0	0.02758
59	MORF_PPP1CC	120	-0.5176	-1.764	0.0077	0.02866
60	HOUSTIS_ROS	22	-0.5459	-1.7571	0.0108	0.03085
61	HUMAN_MITODB_6_2002	278	-0.4684	-1.7489	0.004	0.03356
62	GCM_ANP32B	28	-0.6489	-1.7464	0.0039	0.03377

63	GLYCINE_SERINE_AND_THREONINE_METABOLISM	23	-0.6235	-1.747	0	0.03409
64	MORF_EIF3S2	172	-0.5033	-1.741	0	0.03596
65	SANA_IFNG_ENDOTHELIAL_DN	53	-0.5347	-1.7372	0.0042	0.03695
66	GNF2_RAN	61	-0.5908	-1.7321	0.0154	0.03853
67	PYRIMIDINE_METABOLISM	34	-0.5768	-1.7302	0.0116	0.03885
68	MANALO_HYPOXIA_DN	64	-0.4815	-1.7281	0	0.03909
69	POMEROY_DESMOPLASIC_VS_CLASSIC_MD_UP	21	-0.5348	-1.7248	0	0.04011
70	GNF2_PA2G4	61	-0.5921	-1.719	0.0197	0.04054
71	GCM_CSNK2B	73	-0.5813	-1.7227	0	0.04055
72	MORF_RAD54L	80	-0.5106	-1.7212	0.0113	0.0407
73	POMEROY_MD_TREATMENT_GOOD_VS_POOR_DN	21	-0.559	-1.7199	0.0167	0.04075
74	PENG_GLUTAMINE_DN	212	-0.4627	-1.7168	0.0041	0.04098
75	HDACI_COLON_CUR24HRS_UP	22	-0.5375	-1.7148	0.0039	0.04106
76	AGUIRRE_PANCREAS_CHR19	30	-0.5984	-1.7022	0.0146	0.04569
77	PENG_LEUCINE_DN	122	-0.4756	-1.7023	0.012	0.04624
78	MORF_PRKDC	144	-0.4867	-1.698	0.0081	0.0472
79	UVB_NHEK3_C6	19	-0.5557	-1.6892	0.0038	0.04866
80	BRCA1_OVEREXP_DN	82	-0.5114	-1.6909	0.0114	0.04892
81	MORF_BUB3	215	-0.4741	-1.6931	0.0119	0.04909
82	OXSTRESS_RPE_H2O2HNE_DN	24	-0.5665	-1.6893	0.0184	0.04921
83	MITOCHONDRIA	283	-0.4253	-1.6915	0	0.04928
84	RNA_TRANSCRIPTION_REACTOME	29	-0.5404	-1.686	0.0041	0.04955

10.3 GSEA-Auswertung – Liste der Gensets, die eine signifikante Anreicherung von Genen aufweist, die durch c-Myc herunterreguliert werden

Rang	NAME	SIZE	ES	NES	NOM p-val	FDR q-val
1	GNF2_CD97	22	0.7801	2.1302	0.0000	0.0083
2	CALCINEURIN_NF_AT_SIGNALING	47	0.5810	2.0952	0.0000	0.0063
3	CELL_MOTILITY	64	0.5667	2.0639	0.0000	0.0064
4	GNF2_RAP1B	25	0.7377	2.0202	0.0036	0.0141
5	EGFPATHWAY	22	0.6861	2.0200	0.0000	0.0113
6	BASSO_GERMINAL_CENTER_CD40_UP	50	0.6663	2.0079	0.0000	0.0118
7	GLEEVECPATHWAY	17	0.6647	1.9974	0.0000	0.0142
8	GNF2_ITGB2	31	0.6402	1.9973	0.0000	0.0124
9	INSULINPATHWAY	17	0.6506	1.9630	0.0000	0.0181
10	CARIES_PULP_UP	94	0.5631	1.9598	0.0036	0.0164
11	TCRPATHWAY	34	0.6633	1.9566	0.0000	0.0153
12	GNF2_ITGAL	27	0.7408	1.9564	0.0000	0.0140
13	PASSERINI_ADHESION	22	0.6549	1.9515	0.0000	0.0140
14	YU_CMYC_DN	32	0.6308	1.9433	0.0037	0.0146
15	GNF2_PTPRC	35	0.7020	1.9407	0.0000	0.0149
16	METPATHWAY	24	0.5973	1.9340	0.0000	0.0149
17	BCRPATHWAY	27	0.6167	1.9267	0.0036	0.0175
18	HALMOS_CEBP_UP	22	0.5661	1.9045	0.0000	0.0213
19	GNF2_PECAM1	21	0.7129	1.9011	0.0000	0.0213
20	GNF2_CASP1	45	0.5983	1.9011	0.0000	0.0202
21	ASTON_DEPRESSION_UP	22	0.6419	1.9006	0.0000	0.0197
22	PYK2PATHWAY	23	0.5916	1.8980	0.0036	0.0198
23	GNF2_SELL	26	0.7242	1.8855	0.0000	0.0222
24	PDGFPATHWAY	23	0.6473	1.8788	0.0000	0.0235
25	FCER1PATHWAY	31	0.5723	1.8720	0.0000	0.0243
26	CHIARETTI_T_ALL	121	0.5034	1.8634	0.0000	0.0264
27	AT1RPATHWAY	28	0.5439	1.8593	0.0033	0.0267
28	GNF2_CASP8	16	0.7627	1.8519	0.0036	0.0288
29	BIOPEPTIDESPATHWAY	29	0.5785	1.8491	0.0000	0.0301
30	CELL_ADHESION_MOLECULE_ACTIVITY	53	0.5460	1.8356	0.0140	0.0348
31	INTEGRIN_MEDIATED_CELL_ADHESION_KEGG	59	0.5232	1.8352	0.0000	0.0339

32	DER_IFNA_UP	34	0.5612	1.8275	0.0000	0.0351
33	IGF1PATHWAY	17	0.6266	1.8255	0.0038	0.0347
34	PASSERINI_SIGNAL	181	0.4321	1.8226	0.0000	0.0345
35	GNF2_JAK1	15	0.8085	1.8223	0.0000	0.0336
36	CHIARETTI_T_ALL_DIFF	133	0.4671	1.8149	0.0000	0.0364
37	NAKAJIMA_MCS_UP	34	0.5261	1.8026	0.0000	0.0421
38	GNF2_MCL1	27	0.6223	1.7991	0.0000	0.0435
39	IL1_CORNEA_UP	28	0.5827	1.7990	0.0000	0.0424
40	CELL_SURFACE_RECEPTOR_LINKED_SIGNAL_TRANSDUCTION	50	0.5046	1.7988	0.0038	0.0416
41	IFN_BETA_GLIOMA_UP	34	0.5495	1.7960	0.0000	0.0417
42	DER_IFNG_UP	31	0.5623	1.7956	0.0036	0.0408
43	IFN_GAMMA_UP	18	0.6201	1.7842	0.0038	0.0470
44	AS3_FIBRO_DN	18	0.6219	1.7823	0.0035	0.0465
45	SANA_TNFA_ENDOTHELIAL_UP	30	0.6193	1.7777	0.0000	0.0476
46	KERATINOCYTEPATHWAY	31	0.5514	1.7747	0.0000	0.0477
47	TAKEDA_NUP8_HOXA9_8D_UP	53	0.4701	1.7731	0.0000	0.0475
48	BRENTANI_CELL_ADHESION	53	0.4532	1.7729	0.0000	0.0465
49	FLECHNER_KIDNEY_TRANSPLANT_REJECTION_UP	40	0.5844	1.7716	0.0071	0.0466
50	GNF2_ICAM3	19	0.6327	1.7693	0.0111	0.0472
51	STEMCELL_COMMON_DN	25	0.5291	1.7673	0.0000	0.0472

10.4 Dialign 2-2 – Ausgabedatei des Programmes für das Gen mit der Unigene ID Hs.534318.

In dieser Datei werden konservierte Sequenzbereiche als Großbuchstaben dargestellt und nicht-konservierte Bereiche in Kleinbuchstaben.

>Hs,534318

ctgggaccagtgggAAAGCTGGAGGACAGGCGGCTCTGAAGCACCGAGGAGgggcCTCGGAAGGGG
CGGGGAGG--------------------------------
TCTCCTTGACACAGGTGAGGAGACGGAGAACTGGGGGCACccggcaGCCCAGGCTTCCTGCcgcg
ctctcacgccCTCCAGGGCCATGAGACCCTGGCTCC-----
GAGGTGGCGGGCAGGTTCGGGACCCGAGCGCGGCGGGATcgtgcggttgccggtggtgctgggccCCATG
GAGGCTGCGGTCCAGGAGTTTCGGGTGGGCg----------
GCTAGGCCGAGCTTCTGCAGGCGTGCTGAGCCGCCg----------
CAGTAGGGCCCCAGAGGGGACGAGGAATTAGGcccagccgaacgaggGGGAGCCACCAGAGGCTG
GGCCGGGAGGCACCGCCGCGCTctggGGAGCTGCCAGGGTCGGGGCgc-CCTCCGAGCGG--------
----
CTGCCannnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnntTCCCCGCaggtcgcCCTGCCAGCTTGAGCGGC
GGGCGCGGGTGGGGCaggagga------
CGCGGGGCGGGGTGGGGGGCACAGGGGGCCCTCTCTGCannnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnaGTCGAGGATTAGCGCGTTCGCGGCCGGCGctgCGGGATTAACCCGCGTGGACtgGACG
CCCGGCCCGGGGATTACTGCGCgCTCCCTCC--
CCGACGTATATATTnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnTGCCCCTTCTCCATCGAGCACATCCTCTCCAGCCTGCCcGAGCGGAGCCT
CCCGGCCCGGgccgcctgcCCACCGCAGCCCGCCGGTCGCCAGAGCCCCGCgaAGCCAGAGGAGC
CCGGGGCGCCCGag---
GCTGCGCCCTGCGCCTGCTGCTGCTGCTnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn--
----

>mouse

----------AAATGTGGTGTGAGGGAGATGTTGATGTACACAGGAG----
CCGGGCAAGGGCGTCAGGGaccttctaagacaggagataacaagaaaaataTCTTCTAGTCTCT
AGAAAGAAGGCACAGAATTGGCAGCTCgc----GCCTTTGCTCCCAGCa----------
--
CTCAGGAGACAAGAAGCAGGCGAATCtttatGAGTTGGATGTCAGTCTCAGCTT
CCCATCCTGTAGGTATtctt----------

CTATAGGAGATTCAGACCAGGGCTTCAAAGTGAGCaagagctgtcaaGTAGGGTC
ATCCTGCTGACAGGGTGCTGAGCTCACtggaaatcttcattCCAGGAATCTCCAGTTT
CATGGAGGGAATGAG----------
GGAAGCCAGGGGCAGAGAGGTGGCGGGGACCCTGTGAGTTa---
GAGGCTTCCAAGGTCCTGGCattCCTCCTGGCAGgcgaccttaagcCTGCCttaaaagatata
aaggatgggctcctga----------TCCCTTCcttt---
CCCTCCAGCTCTAGACGTGGCAGAGTGCAGGACtccctgctcccagCGCAGTCCCC
GGTACCCAGACGAAGGGGCTTGCCCTCCtccatg----------
GCCGGGGATTAGCTCTCTGAGGACTGGCGa--
CGGGATTAGCACGCGCGAACgcGGCGCCCGGTCTGGGGATTACCGCAC-
CTCCCACCtcCCTATGTATATATTccgcctcttgcgccgcccgcccgcgggcatggcgactgcaggcag
cgcggccagccgcagggaccccgggcgaccc-------
TGCCCTTTCTCCATCGAGCACATCCTCTCCAGCCTGCCgGAGCGCAGGCCCG
CGACGCGG---------
CCACCGCAGCCCGTCGGTGGTCGGAACCCCGCcgAGCTAGACGAACCCGAG
GCGCCCGtccccGCTGCTCCTTGCGCCTGCTGCTGCTGCTgcaacccgcgcgctgcgaccc
gcggaaccccggagacatcgtccgggccgggtgagtgcgcgcggaggggcaagactaggcggggatctcgcggggcg
gcct

10.5 Experimentenliste

Experiment 1

cDNA Mikroarrayversuch zur Untersuchung der c-Myc abhängigen Genexpression in T-Lymphozyten von transgenen Mäusen (s. 3.4.1). Der Versuch wurde durchgeführt von Tobias Otto (AG Eilers) am IMT in Marburg.

Experiment 2

siRNA Screen mit dem der Einfluss verschiedener Kinasen auf die Stabilität von c-Myc getestet werden soll. Hierzu wurde die aus 779 siRNA-Pools bestehende *siARRAY Human Protein Kinase Library* der Firma Dharmacon verwendet. Der Versuch wurde durchgeführt von Theresia Kress (AG Eilers) am IMT in Marburg.

Experiment 3

shRNA Screen mit dem der Einfluss verschiedener Kinasen auf die Stabilität von c-Myc getestet werden soll. Hierzu wurde eine aus 638 shRNA Pools bestehende Bibliothek des *Netherlands Cancer Institute* (NKI, Amsterdam) (Bernards, Brummelkamp et al. 2006) verwendet. (s. 6.3.1). Der Versuch wurde durchgeführt von Theresia Kress (AG Eilers) am IMT in Marburg.

Experiment 4

Testversuch zur Flachbildkorrektur, durchgeführt am IMT Marburg

11 Lebenslauf

Birgit Samans
geb. am 29.10.1966 im Aachen
Staatsangehörigkeit: deutsch

07/1972-07/1976	Katholische Grundschule Eschweiler Stadtmitte
08/1976-07/1985	Liebfrauenschule Eschweiler, Gymnasium
05/1985 – 08/1985	Volontariat in einem Kibbutz in Israel
10/1985 – 12/1985	Aupair Aufenthalt in London (England)
02/1986 – 08/1086	Volontariat in einem Kibbutz in Israel
10/1986 – 07/1994	Studium der Biologie an der Justus Liebig Universität in Giessen Prüfungsfächer: Genetik, Zoologie, Zellbiologie, Tierphysiologie
02/1993 – 07/1994	Diplomarbeit am Institut für medizinische Mikrobiologie des Klinikums der JLU Gießen bei Herrn Prof. Dr. Chakraborty Thema: Molekulare Untersuchungen zur Regulation der Virulenz bei *Listeria monocytogenes* durch den Regulator PrfA
05/1994 – 06/1994	Studentische Hilfskraft am Institut für medizinische Mikrobiologie der JLU Gießen
09/1994 – 12/1994	Wissenschaftliche Hilfskraft am Institut für medizinische Mikrobiologie der JLU Gießen
01/1995 – 05/1995	Honorarverträge bei einzelnen Messen, Messegesellschaft Frankfurt
05/1995 – 07/1996	Pädagogische Mitarbeiterin im Kinder- und Jugendwohnheim St. Stephanus des Caritas-Verbandes in Giessen
07/1996 – 04/1998	Erziehungszeit (ein Sohn, geb. 07.08.1996)
05/1998 – 08/1998	Randstad Zeitarbeit GmbH
09/1998 – 06/2000	Marketingassistentin bei der ScheBo Biotech AG
07/2000 – 01/2002	Informationsbrokerin/Kundenbetreuerin bei Scitari Informationsdienste
05/2002 – 06/2002	Wissenschaftliche Hilfskraft am Institut für Agrarsoziologie und Beratungswesen, Justus Liebig Universität Gießen

07/2002 – 05/2004	Wissenschaftliche Mitarbeiterin am Institut für Medizinische Biometrie und Epidemiologie, Philipps-Universität Marburg
Seit 06/2004	Wissenschaftliche Mitarbeiterin am Institut für Molekularbiologie und Tumorforschung, Philipps-Universität Marburg

Teile dieser Arbeit sind veröffentlicht in:

Kleinschmidt MA, Streubel G, Samans B, Krause M, Bauer UM. The protein arginine methyltransferases CARM1 and PRMT1 cooperate in gene regulation. Nucleic Acids Res. 2008 Jun; 36(10):3202-13.

Hausmann A, Samans B, Lill R, Mühlenhoff U. Cellular and mitochondrial remodeling upon defects in iron-sulfur protein biogenesis. J Biol Chem. 2008 Mar 28; 283(13):8318-30.

Schulte JH, Horn S, Otto T, Samans B, Heukamp LC, Eilers UC, Krause M, Astrahantseff K, Klein-Hitpass L, Buettner R, Schramm A, Christiansen H, Eilers M, Eggert A, Berwanger B. MYCN regulates oncogenic MicroRNAs in neuroblastoma. Int J Cancer. 2008 Feb 1;122(3):699-704.

Osterloh L, von Eyss B, Schmit F, Rein L, Hübner D, Samans B, Hauser S, Gaubatz S. The human synMuv-like protein LIN-9 is required for transcription of G2/M genes and for entry into mitosis. EMBO J. 2007 Jan 10; 26(1):144-57.

Nowak K, Kerl K, Fehr D, Kramps C, Gessner C, Killmer K, Samans B, Berwanger B, Christiansen H, Lutz W. BMI1 is a target gene of E2F-1 and is strongly expressed in primary neuroblastomas. Nucleic Acids Res. 2006 Mar 31; 34(6):1745-54.

Zirn B, Samans B, Wittmann S, Pietsch T, Leuschner I, Graf N, Gessler M. Target genes of the WNT/beta-catenin pathway in Wilms tumors. Genes Chromosomes Cancer. 2006 Jun; 45(6):565-74.

Slater EP, Diehl SM, Langer P, Samans B, Ramaswamy A, Zielke A, Bartsch DK. Analysis by cDNA microarrays of gene expression patterns of human adrenocortical tumors. Eur J Endocrinol. 2006 Apr; 154(4):587-98.

Gebhardt A, Frye M, Herold S, Benitah SA, Braun K, Samans B, Watt FM, Elsässer HP, Eilers M. Myc regulates keratinocyte adhesion and differentiation via complex formation with Miz1. J Cell Biol. 2006 Jan 2; 172(1):139-49.

Zirn B, Hartmann O, Samans B, Krause M, Wittmann S, Mertens F, Graf N, Eilers M, Gessler M. Expression profiling of Wilms tumors reveals new candidate genes for different clinical parameters. Int J Cancer. 2006 Apr 15; 118(8):1954-62.

Zirn B, Samans B, Spangenberg C, Graf N, Eilers M, Gessler M. All-trans retinoic acid treatment of Wilms tumor cells reverses expression of genes associated with high risk and relapse in vivo. Oncogene. 2005 Aug 4;24(33):5246-51.

Göke R, Barth P, Schmidt A, Samans B, Lankat-Buttgereit B. Programmed cell death protein 4 suppresses CDK1/cdc2 via induction of p21(Waf1/Cip1). Am J Physiol Cell Physiol. 2004 Dec; 287(6):C1541-6.

12 Akademische Lehrer

Meine akademischen Lehrer waren die Herren und Damen

In Giessen

Anders, Askani, Bentrup, Blecher, Brändle, Chakraborty, Clauß, Cleffmann, Hegemann, Hobom, Jäger, Jauker, Löb, Nowak, Renkawitz, Ringe, Schipp, Schoner, Schulte, Schwartz, Seibt, Steubing, Werding

In Marburg

Eilers

13 Danksagung

Ich möchte mich herzlich bei allen bedanken, die zum Gelingen dieser Arbeit beigetragen haben.

Herrn Prof. Dr. Martin Eilers, dass er mir die Möglichkeit zur Anfertigung einer Dissertation in seiner Arbeitsgruppe gegeben hat, seiner Betreuung und seinen wertvollen Anregungen.

Herrn Prof. Dr. Schäfer für seine statistische Beratung.

Allen Mitarbeitern der AG Eilers und der Microarray Unit für die gute Zusammenarbeit und für das angenehme Arbeitsklima.

Mona Meyer und Theresia Kress fürs Korrekturlesen und ihre konstruktiven Kritiken.

Matthias Danne für seine unerschöpfliche Geduld und Unterstützung.

Und schließlich möchte ich mich ganz besonders bei meinem Sohn Natnael bedanken, der mir die vielen am Computer verbrachten Stunden verzeihen möge. Sein Dasein und seine Lebensfreude haben mir immer wieder die Motivation und Energie zum Weiterführen dieser Arbeit gegeben.

14 Ehrenwörtliche Erklärung

„Ich erkläre ehrenwörtlich, dass ich die dem Fachbereich Medizin Marburg zur Promotionsprüfung eingereichte Arbeit mit dem Titel „Statistische Auswertung und Interpretation von hochdimensionalen molekularbiologischen Datensätzen" am Institut für Molekularbiologie und Tumorforschung (IMT) der Philipps-Universität Marburg unter Leitung von Prof. Dr. Martin Eilers ohne sonstige Hilfe selbst durchgeführt und bei der Abfassung der Arbeit keine anderen als die in der Dissertation angeführten Hilfsmittel benutzt habe. Ich habe bisher an keinem in- und ausländischen medizinischen Fachbereich ein Gesuch um Zulassung zur Promotion eingereicht, noch die vorliegende oder eine andere Arbeit als Dissertation vorgelegt.